Dispositifs Infocommunicationnels : des traces numériques d'usage aux données d'analyse

Jean-Marc Francony

UNIVERSITÉ Grenoble Alpes

Dispositifs info-communicationnels :
des traces numériques d'usage aux données d'analyse

Mémoire d'Habilitation à Diriger des Recherches

Jean-Marc FRANCONY

Soutenue à Grenoble, le 13-06-2017

Membres du Jury :

Pr. Laurence BALICCO	Université Grenoble-Alpes (GRESEC)
Pr. Madjid IHADJADENE	Université de Paris VIII (PARAGRAPHE)
Pr. Pascal MARCHAND	Université de Toulouse (LERASS)
Pr. Isabelle PAILLIART	Université Grenoble-Alpes (GRESEC) - Garante
Pr. Françoise PAQUIENSEGUY	Institut d'Étude Politique Lyon (ELICO)

À Aline
Et toute la tribu,

Remerciements

Je remercie vivement les professeures et professeurs, Laurence BALICCO, Madjid IHADJADENE, Pascal MARCHAND, Isabelle PAILLIART, Françoise PAQUIEN-SEGUY qui m'ont fait l'honneur d'être rapporteurs de ce mémoire d'habilitation à diriger les recherches et de participer à mon jury de soutenance.

Je remercie plus particulièrement Isabelle PAILLIART Professeure en Sciences de l'Information et de la Communication (SIC), directrice du GRESEC de l'Université Grenoble Alpes (UGA) qui a accepté d'être garante de mon travail. Sa connaissance de la discipline et ses conseils m'ont été d'un précieux secours tout au long du processus d'écriture. Son soutien et ses critiques toujours constructives me permettent de me présenter devant vous aujourd'hui.

Le travail présenté dans ce mémoire est par nature collectif. Mes plus sincères remerciements vont à toutes celles et tous ceux avec qui j'ai été associé dans la conduite de projets et avec qui j'ai partagé non seulement le plaisir d'une recherche collective mais aussi celui d'éclairages différents qui existent au sein de notre discipline comme dans d'autres.

Parmi ces personnes, je tiens à distinguer tout particulièrement Gilbert EYMARD (Maître de Conférences SIC - GRESEC) trop tôt disparu, et Françoise PAPA (Maître de Conférences SIC - PACTE). J'ai pu m'enrichir à leur contact de leur sensibilité et de leur approche des Sciences de l'information et de la communication autant que de leurs goûts et de leurs connaissances des Sciences humaines et sociales. Au fil des ans nous avons su construire notre langage commun et mettre à profit nos différences autant que nos aspirations partagées pour mener de concert un travail qui s'est toujours porté sur des questions émergentes et complexes

Je tiens également à remercier, mes collègues chercheurs de l'UMR PACTE, sociologues, politologues, géographes et statisticiens, avec lesquels l'interdisciplinarité est une évidence et un mode d'existence scientifique au quotidien.

Je tiens enfin à remercier mes collègues Evelyne MOUNIER (Maître de Conférences en SIC) et Aude INAUDI (Maître de Conférences en SIC) du GRESEC pour leurs relectures et leur indéfectible solidarité amicale qui fonde l'équipe du Master *Métiers du Livre et de l'Édition* de l'UGA. Porter ensemble une formation nous conduits plus que jamais à renforcer une dynamique scientifique disciplinaire et à susciter des projets entre nos laboratoires. J'associe également Édith LAVIEC et Matthieu MEYNET (Doctorants au GRESEC) qui sont les premiers témoins de cette continuité.

Enfin, je n'oublie pas tous les collègues et administratifs des composantes de l'Université, tout particulièrement ceux de l'UMR PACTE et de l'UFR Sciences de l'Homme et de la Société dont je partage le quotidien et sans qui nos missions d'enseignant-chercheur seraient impossibles.

Sommaire

Table des figures

Préambule

La préparation d'une Habilitation à Diriger les Recherche convoque individuellement le chercheur et le questionne sur son parcours et ses productions scientifiques[1]. La pratique de la recherche interdisciplinaire et collective que je soutiens, m'a amené à hybrider mes réflexions et à nouer en permanence des collaborations dont attestent mes publications. Le mémoire que je présente se donne comme objectifs de décrire le programme scientifique que je poursuis, d'asseoir sa légitimité au sein des Sciences de l'Information et de la Communication (SIC), et de mettre en lumière les questionnements épistémologiques et théoriques que l'évolution des technologies numériques de l'information-communication maintient ouverts dans cette discipline.

Le domaine auquel mes travaux se consacrent est celui des médiations numériques impliquant des *sujets*[2] humains. Le terme de *médiation* insiste sur l'existence de l'intermédiaire, le médiateur, qui s'intercale dans les relations interpersonnelles qu'il canalise. Ce médiateur apparait tantôt dans sa fonction générale de connecteur, dans une acception neutralisée, tantôt dans sa fonction de passeur, suivant un sens plus investi (Jeanneret, 2009). L'évolution des technologies et des pratiques sociales centrées sur *le numérique*[3] incite à de nouvelles approches théoriques et pratiques. La nouveauté n'est pas seulement liée à l'évolution de l'objet de recherche. Elle affecte aussi l'instrumentation que peuvent mobiliser les chercheurs dans leurs études, établissant ainsi une continuité numérique entre l'observé et l'observateur. Ce continuum mobilise de nouvelles techniques dont la mise en œuvre interroge les méthodes classiques jusque-là utilisées dans le domaine des Sciences humaines. Les chercheurs, notamment en SIC, sont de ce fait confrontés à un virage numérique ou *digital turn* voire d'un *turning point*, qu'il leur faut assurer afin de supporter un programme d'humanités numériques ou *Digital Humanities*[4] (Berry, 2011), (Abbott, 2009).

D'un point de vue théorique, la notion de *dispositifs info-communicationnel*[5] me parait s'imposer pour décrire la complexité des interactions et des configurations d'acteurs mobilisés dans les médiations numériques. D'un point de vue pratique, les conditions de connexion permanente en tout temps et en tout lieu ont installé des pratiques qu'il devient difficile d'associer spécifiquement à des rythmes de vie ou à des espaces spécialisés. L'observation des usages et des pratiques sociales requiert en conséquence de nouvelles méthodes qui permettent un suivi continu des sujets et de leurs interactions numériques.

---

[1] Voir le document annexe : *sélection de publications*. Dans la suite du document, les références personnelles sont associées aux notes chronologiques Ref.x.

[2] Le terme est utilisé ici dans le sens qui lui est donné dans le domaine de l'expérimentation en psychologie.

[3] J'utiliserai, bien que ce soit un abus de langage *cet adjectif substantivé* (Moatti, 2012) comme équivalent de l'expression de technologies numériques qui sont par essence des technologies info-communicationnelles (TIC).

[4] On reste pour l'instant sur une définition intuitive et dans un sens large, définition qui sera précisée dans le chapitre 6.

[5] La notion fait l'objet d'un long développement au chapitre 1.

La *traçabilité* est essentiellement envisagée en référence aux contraintes légales d'une production industrielle qui imposent le suivi des chaines de transformation des ingrédients puis de la distribution du produit fini. Dans les termes courants, la traçabilité est un engagement responsable du producteur auprès du consommateur. Dans le contexte numérique, une forme équivalente d'engagement existe pour les prestataires de services qui doivent garantir l'accomplissement de prestations contractuelles (commerce en ligne, hébergement, etc.). Pour ces derniers, la traçabilité doit aussi permettre d'assumer leur statut d'éditeurs des contenus qu'ils publient[6]. Mais dans la majorité des cas, le principe de traçabilité se renverse au profit d'une responsabilité de mise en œuvre, c'est-à-dire d'usage. C'est la raison pour laquelle, le *traçage*[7] numérique des individus s'est d'abord développé dans la continuité d'une logique d'*accountability*[8] propre à garantir la sécurité fonctionnelle des systèmes informatiques. Mais si la *traçabilité* est une nécessité technique et juridique qui impose la constitution de *traces d'usage*[9] authentifiées, ces dernières se développent désormais bien au-delà de ce cadre, dans une perspective de création de valeur. La transformation de traces personnalisées d'usage en données numériques est devenue une réalité industrielle et marchande très prégnante.

L'effet d'accélération qui accompagne le renforcement constant des technologies numériques provoque de multiples décalages au sein des sociétés contemporaines et démocratiques entre les opportunités techniques et les cadres de régulation de l'action. Ces décalages interrogent le devenir des sociétés (de la connaissance ?) et de la démocratie (délibérative ?). La mise en données du monde ou *datafication*[10] est l'un des éléments du débat mais pas le seul. Il fait porter sur les données numériques un regard ambivalent fait de craintes et d'espoirs hypothétiques.

Les *humanités numériques*[11] trouvent une légitimité dans ces débats auxquels elles sont conviées. La *datafication* représente une opportunité pour les chercheurs en SHS de suivre les traces traduites en données comme moyen d'investigation de l'espace numérique. La circulation de ces données dans l'espace d'accès public qu'est le Web permet de formuler des hypothèses sur les processus info-communicationnels, sur les logiques d'acteurs ainsi que les chaines de valeurs qui tissent conjointement le marché des données dans ses régulations complexes. Plus en profondeur, étudier ces gisements de données dans ce qu'ils décrivent les singularités d'une personne que ce soit dans

---

[6] C'est le sens de la jurisprudence qui s'applique dans ce secteur.

[7] J'utiliserai ce terme dans une double référence : celle de l'informatique où cette notion évoque les systèmes de mise au point (trace) en suivant l'exécution pas à pas d'un programme informatique ; en filant la métaphore minière de la fouille des données (cf. chapitre 3), celle associée «*au creusement d'une galerie dans un gisement minier, en vue de sa reconnaissance ou de son exploitation ultérieure*». http://www.cnrtl.fr/definition/traçage

[8] Voir chapitre 3 pour sa définition

[9] Dans tout le mémoire, j'ai fait le choix d'écrire au singulier le mot "usage" pour les expressions *traces d'usage*, *données d'usage*. Ce singulier renvoie à l'individualité de l'usager qui laisse des traces. La singularisation des traces est également un objectif dans la mise en œuvre du traçage numérique à des fins analytiques ou de valorisation (personnalisation ou segmentation). C'est enfin en référence aux travaux des archéologues préhistoriens qui dès le début du XXe siècle évoquent les *traces d'usage* sur les outils cf. chapitre 3.

[10] Voir chapitre 5 pour sa définition

[11] Sujet abordé dans le chapitre 6.

ses pratiques info-communicationnelles, dans son activité, mais aussi dans ce qu'ils permettent de calculer et d'inférer introduit nécessairement des hypothèses corollaires sur les objets et la nature des interactions. Au travers de ces hypothèses, c'est la nature des interactions sociales et de leurs objets qui deviennent atteignables à une échelle inégalée jusque-là.

Le projet de mémoire que je soutiens s'inscrit dans le double mouvement d'une recherche sur le numérique et au moyen du numérique. L'emploi d'outils et de méthodes numériques a été une manière de mobiliser mes compétences techniques dans le champ des SIC et d'y évoluer[12]. Cette orientation s'inscrit dans la logique des *Digital Methods* que promeut Richard Rogers (Rogers, 2009, 2010, 2015). Elle s'accompagne d'un intérêt certain pour les questions méthodologiques que le numérique exige.

Ce mémoire est l'expression d'un engagement affirmé dans les Sciences de l'Information et de Communication. Sa rédaction est guidée par l'actualité du questionnement méthodologique sur la prise en compte du numérique et des données dans la conduite des travaux en Sciences sociales et en SIC. Le point de départ de ce questionnement porte sur l'observation numérique de dispositifs info-communicationnels au travers des traces numériques qu'ils engendrent sur le Web. Appuyant ma réflexion sur cette problématique ce sont les conditions de production du savoir dans le contexte numérique qui sont interrogées et de leurs conséquences sur l'évolution de notre discipline.

[12] Le chapitre introductif retrace l'évolution de mon parcours scientifique.

Introduction

In a few years,
men will be able to communicate
more effectively through a machine than face to face.
That is a rather startling thing to say,
but it is our conclusion.
(Licklider, Taylor, 1968)

L'objectif de ce chapitre introductif est de montrer comment mes approches disciplinaires et scientifiques (§3) se sont élaborées dans un cheminement personnel qui m'a conduit de l'Informatique aux Sciences de l'information et de la communication (§1, §2).

1. Une inscription en Sciences humaines et sociales

Mon parcours a d'abord été celui de l'ingénierie informatique avant de devenir spécifique aux domaines d'application des Sciences Humaines et Sociales (SHS) à l'issue d'un DEA puis d'un doctorat d'informatique appliquée aux Sciences sociales conduit au CRISS[13] sous la responsabilité du professeur Jacques Rouault (71e section).

Cette orientation en SHS s'est construite par attrait pour des sujets tels que l'intelligence artificielle et la représentation des connaissances, mais dont la compréhension réclamait des ponts vers d'autres disciplines. C'est cette connaissance plurielle que je suis venu chercher au CRISS, alors en pointe dans les domaines du traitement automatique de la langue naturelle écrite (TALN) et dans les systèmes experts. Il en découlait une richesse et une diversité de collaborations scientifiques, notamment avec le laboratoire de psychologie et d'ergonomie de l'INRIA, tout juste délocalisé à Grenoble, et animé par André Bisseret (psychologue ergonome). L'expertise du CRISS sur ces différents sujets s'est construite dans la conduite d'un programme scientifique rigoureux et méthodique piloté par J. Rouault, ainsi que dans la participation de l'équipe aux appels d'offres de la communauté européenne (programmes Esprit), selon une perspective d'ingénierie et de transfert industriel. Cette double orientation répondait à mes aspirations. J'ai donc investi cette voie.

1.1. Interfaces et interactions multimodales

Dans les années 1980, les dispositifs informatiques interactifs atteignent une telle complexité qu'il n'est plus possible d'envisager rationnellement les développements de l'industrie logicielle sans une meilleure connaissance des problématiques de l'interactivité. La réponse apportée à ce problème industriel est double : d'une part, favoriser la séparation entre les différents composants

[13] Centre de Recherche en Informatique appliquée au Sciences Sociales – Université Pierre Mendès France, Grenoble 2

fonctionnels réalisant la partie interactive et la partie applicative spécifique de la machine ; d'autre part, constituer une expertise dans la conception des systèmes interactifs. En 1990, Joëlle Coutaz (Coutaz, 1990) professeure d'informatique estime que le premier objectif ressort du domaine des interfaces homme-ordinateur alors que le second concerne le domaine des interactions homme-machine.

Ainsi, la question des interfaces relève clairement de la discipline informatique dont elle constitue un sous-domaine, alors que celle des interactions lui est extérieure dans la mesure où le concept de machine renvoie à une définition plus vaste que les seules réalisations informatiques. Au début des années 1990, l'interaction homme-ordinateur se situe à l'intersection de ces deux points de vue, ce qui a permis d'ouvrir un nouveau champ de recherches et de collaborations interdisciplinaires.

À cette époque, la métaphore du dialogue interpersonnel rapportée au couple homme-machine est considérée comme l'aboutissement de la forme naturelle de l'interaction. Dialoguer signifie dans ce cas interagir pour contribuer au déroulement de l'activité suivant un fil naturel mais aussi interagir sur les termes et le procès de l'interaction. Dans ce contexte, l'approche multi-modes (écrit/oral, gestuel) du dialogue homme-ordinateur s'est imposée comme un degré supplémentaire de liberté accordée aux individus dans le choix des modalités (langue naturelle, etc.) de l'interaction. Ainsi, l'intervention d'un individu, comme celle de l'interface peut mobiliser différentes modalités, joindre par exemple le geste à la parole.

C'est dans ce contexte scientifique qui attribuait à l'utilisateur final (*End User*) un rôle clé dans le processus d'élaboration logiciel, non seulement comme acteur de celui-ci mais également comme objet de conceptualisation, qu'un projet de thèse portant sur la modélisation du dialogue homme-machine multimodal[14] m'a été proposé (Ref.5). Ce sujet s'inscrivait dans le domaine des interactions homme-machine et mettait l'accent sur une nécessaire interdisciplinarité avec les domaines des SHS.

Cette thèse a débuté au moment où le CRISS s'engageait dans le projet Européen Esprit : MMI2 (*Multi-Modal Interface for Man-Machine Interaction*); l'objectif de la collaboration était de réaliser un démonstrateur[15] intégrant un ensemble de modalités se rapportant aux langues naturelles écrites (français, anglais, espagnol) et à des langages d'expressions gestuelles et graphiques.

L'engagement scientifique du CRISS dans le contrat MMI2 visait plus particulièrement la modalité de la langue naturelle écrite pour le français. Ce projet représentait l'opportunité de finaliser et d'intégrer les modules de traitements (TALN[16]) jusqu'alors disjoints, dans une chaine unifiée de traitements pour en éprouver la viabilité et la robustesse. Ce sujet de thèse s'articulait avec d'autres axes de recherches du laboratoire portant sur la génération automatique de réponses écrites ou sur l'interface de dialogue.

La proposition de thèse qui m'a été faite comportait une partie exploratoire importante ouverte sur ce nouvel objet que constituait l'interaction dialogique multimodale. Compte tenu de l'historique

[14] Articulant différents modes et modalités

[15] Présentation du démonstrateur (MMI2, 1993)

[16] Traitement Automatique des Langues Naturelles.

du laboratoire, ce sujet comportait une hypothèse forte portant sur le rôle privilégié de la langue naturelle dans la conduite d'une interaction multimodale.

J'ai débuté ces études doctorales dans la continuité d'une pensée ingénieure que l'on peut qualifier de cybernétique (Miege, 1995) visant à la constitution d'un système parfait : le couple homme-ordinateur. Ce n'est qu'au fil de ma thèse qu'une approche différente s'est élaborée, face à un questionnement favorisé par le contexte scientifique interdisciplinaire du laboratoire.

1.2. Les enjeux interdisciplinaires de l'interaction homme-ordinateur

Dans la perspective historique des travaux sur l'interaction homme-ordinateur, la psychologie et l'ergonomie cognitive ont été considérées du point de vue de l'informatique comme des disciplines susceptibles d'apporter des modèles, des méthodes et des éléments pratiques nécessaires à la conception d'interfaces avancées (Coutaz, 1990). Ces apports ont été conséquents, notamment dans l'évaluation des interfaces ainsi que dans la mise en œuvre de la démarche expérimentale pour laquelle ces disciplines disposent de cadres méthodologiques robustes. Alors que celles-ci cherchent à mettre à jour des modèles cognitifs qui dépassent de loin la question du poste de travail, la collaboration interdisciplinaire s'est principalement concentrée sur le couple homme-ordinateur et a reposé sur le paradigme de *performance*[17] associé à l'*utilisabilité*[18] des dispositifs informatisés (Denis, 2009). De fait, ce paradigme convenait (et convient encore) à des situations où l'activité visée par l'utilisateur est finalisée, et correspond à un but qui peut se réaliser par une séquence opératoire de l'application (ou noyau fonctionnel). Cette conception s'appuie sur l'hypothèse selon laquelle l'utilisateur dispose d'une compréhension rationnelle de son activité, mais aussi de son instrumentation (application) et de la situation d'interaction. Ces conditions sont évidemment celles de situations de médiation les plus favorables, où le rôle de l'interface est celui « d'un médiateur » en capacité d'interpréter l'intention performative de l'utilisateur et de le conduire, via une représentation de tâches à assurer, en fonction des fonctionnalités du système (interface + application).

Dans un laboratoire tel que le CRISS, historiquement ouvert à d'autres disciplines comme la linguistique (suivant des présupposés structuralistes) ou les Sciences de l'information et de la communication, cette compréhension fonctionnaliste de l'interaction ne pouvait être pleinement satisfaisante. En revanche, étudier les situations d'échec lors de médiations instrumentales ou bien questionner les formes du dialogue interpersonnel en situation de résolution de problème, correspondaient davantage aux perspectives du laboratoire.

Une telle approche a soulevé de nombreuses questions épistémologiques et engendré de vives discussions, portant sur : les emprunts disciplinaires, la nature des corpus à constituer et des modèles d'analyse (principalement en linguistique), l'ancrage naturel relatif aux interactions et au dialogue humain, la place de la langue naturelle dans les dispositifs d'interaction.

---

[17] La performance fait ici référence à l'optimisation instrumentale dans l'accomplissement d'une tâche. Dans la littérature informatique, de très nombreuses références anglaises associent *performance* et *usability*.

[18] En traduction du concept d'*usability* qui comporte les deux facettes : de prédisposition à l'utilisation, que l'on peut lire comme facilité ; d'utilisation dans la durée qui donne parfois la traduction *usabilité* au sens d'user.

Dans le même temps, un autre espace de confrontation s'était ouvert sur les enjeux des interfaces et de l'interaction dialogique. En effet, le recentrage sur l'utilisateur engagé dans une activité informatisée est aussi très présent à cette époque dans les problématiques des Sciences de l'information et de la communication, notamment vis-à-vis de la recherche informationnelle et documentaire. Cette proximité avec les problématiques de l'interaction homme-machine ne pouvait que favoriser l'émergence, d'une part, de questionnements sur les médiations des activités info-communicationnelles et, d'autre part, de travaux liés aux interfaces des systèmes documentaires. Dans ce contexte, les limites du modèle *Question-Réponse* permettaient de reposer la question de l'interaction informationnelle avec les systèmes documentaires[19]. L'ouverture à d'autres modalités revenait également à questionner la légitimité de la langue naturelle écrite comme modalité privilégiée de l'interaction. L'enjeu pour ce domaine des Sciences de l'information et de la communication se situe au-delà d'un modèle de la tâche à effectuer. Il s'agissait d'intégrer : un modèle de l'usager, une expertise technique sur les langages documentaires et sur les médiations informationnelles. Ces apports correspondent à des enjeux de positionnement scientifiques et politiques. Ceux-ci concernent alors, non seulement le rapport interne avec les sciences de la communication et la proximité avec l'Informatique, mais ils touchent également aux enjeux de légitimité dans la représentation des métiers de la documentation et de la capacité d'une discipline scientifique à en développer une vision prospective, ancrée sur le terrain professionnel.

Ces différents débats m'ont sensibilisé aux difficultés et à l'intérêt d'une approche pluridisciplinaire. Ils m'ont appris le temps nécessaire à l'indispensable naissance d'un langage commun et de représentations partagées. Ils m'ont fait également prendre conscience des enjeux disciplinaires qui se nouent autour de sujets émergents.

Ce parcours de troisième cycle au sein d'une équipe pluridisciplinaire m'a conduit à changer résolument de point de vue et à fixer mon intérêt non plus sur l'informatique en tant qu'objet mais en tant que moyen. En effet, la finalité opérationnelle des modèles est une contrainte très forte à laquelle les approches disciplinaires ne sont pas toujours sensibles ni à même de répondre. Par ailleurs, l'implémentation informatique de modèles issus des théories constitue une mise à l'épreuve formelle redoutable de celles-ci. Dans le cadre de ma thèse, j'ai été confronté aux difficultés d'identifier des modèles et des travaux susceptibles d'apporter un ancrage théorique à la réalisation d'interface de dialogue. Ainsi, le changement de perspective était imposé par le fait de participer à l'évolution scientifique d'un domaine comme celui des interfaces homme-ordinateur ou, de manière plus générale, de contribuer à l'évolution de la prise en compte des modèles de l'utilisateur dans les dispositifs info-communicationnels.

C'est donc la question de l'interaction Homme-machine qui m'a permis de me confronter, dans le contexte des technologies numériques, aux problématiques de l'information et de la communication *médiatisée*[20]. Depuis cette époque et au bénéfice d'un poste de Maître de Conférences en Sciences de l'information et de la communication, ce changement de perspective

[19] Voir à ce sujet l'article d'Hubert Fondin, posant en 2001, la question du choix de paradigme positiviste/ subjectiviste dans les Sciences de l'information (Fondin, 2001).

[20] Je préciserai ce terme dans le paragraphe 3 de cette introduction.

s'est confirmé comme ancrage disciplinaire et comme cadre pour le développement de collaborations avec d'autres champs disciplinaires afin d'aborder des situations d'activités complexes et des problématiques globales. Depuis, à mes yeux, l'interaction homme-machine s'inscrit davantage comme moyen explicite d'aborder des phénomènes sociaux que les dispositifs info-communicationnels soutiennent et dévoilent. Ainsi, comme l'écrit Serge Proulx : « *La technologie devient une dimension de l'écologie humaine et sociale parmi d'autres.* » (Proulx, 2015)

1.3. Le recueil de corpus en langue naturelle écrite

Capter et enregistrer les traces de l'interaction est apparu très tôt dans mes travaux comme un passage obligé. En effet, comment envisager l'interaction dialogique homme-machine alors qu'aucun dispositif informatique ne permet de le supporter ? Le recours à la simulation expérimentale s'impose. Mais alors, comment reconstituer une situation quasi-naturelle (écologique) d'utilisation ? Et quels observables capturer ?

Répondant à un appel d'offre du CNRS pour la constitution de corpus de dialogue homme-machine en langue naturelle (Ref.1), j'ai réalisé un système informatique baptisé *SOPHOCLE* (Ref.2-3) permettant de mettre en œuvre un protocole expérimental de type "magicien d'Oz" à distance, afin de simuler[21] le fonctionnement d'un système d'information dialoguant en ligne et d'en recueillir les traces d'interactions engendrées. L'expérimentation avait pour objectif d'établir et de contrôler un cadre d'observation pour une double pratique informationnelle et communicationnelle. On peut alors parler d'*observation expérimentale*. L'objectif principal poursuivi était d'examiner comment les usagers expriment leurs recherches d'information quand ils s'adressent en langage libre à une machine. L'illusion qu'un dialogue en langue naturelle pouvait alors se nouer était une hypothèse très plausible, et ce n'était pas nouveau, puisqu'il y avait déjà eu le système Eliza (1964-1966) au MIT qui simulait le dialogue avec un psychologue, système qui a démontré que les usagers tendent à personnaliser leur rapport à la machine (Weizenbaum, 1966).

Les sessions recueillies[22] expérimentalement à l'aide de SOPHOCLE comportaient non seulement la séquence horodatée des tours de parole (actes illocutoires) et des écrits produits, mais aussi la ligne temporelle des micro-événements relatifs à l'activation des dispositifs de saisie (clavier et souris). À l'époque, seules les traces dialogiques textuelles ont été retenues et valorisées selon l'engagement contractuel.

L'exploration des micro-événements des enregistrements réalisés, n'a pas été approfondie. Avec le recul, cette absence d'intérêt et de problématiques associées sont aussi à considérer du point de vue théorique et épistémologique. L'approche du sujet énonciateur portée par la théorie linguistique (notamment de l'analyse structuraliste) était dominante. Elle instaurait l'interaction en tant que cadre d'énonciation et lui subordonnait la finalité de l'action. SOPHOCLE assurait le cœur

[21] Le propre de ce protocole réside dans la mise en œuvre simulée des comportements du dispositif de médiation. Dans le cas présent, le comparse produit les réponses de l'interface en s'aidant d'un système documentaire. L'illusion est celle d'une capacité dialogique en langue naturelle écrite. Les attitudes et comportements du comparse sont régis par un ensemble de règles.

[22] Une session est associée à une séquence d'interactions continue pour un même utilisateur.

communiquant d'un dispositif expérimental qu'il fallait construire par ajouts d'éléments externes suivant la situation et le protocole d'observation. L'écologie situationnelle provenait de la conduite *in situ* des observations sur les lieux d'usage ordinaire d'un dispositif informationnel et de l'ignorance qu'avaient les sujets durant l'observation d'être face à un système simulé. SOPHOCLE n'intégrait que faiblement les ressources externes utilisées ce qui ne permettait pas d'enrichir l'enregistrement des échanges de métadonnées contextuelles liées à leur utilisation. La pauvreté de la trace numérique qui en découlait a réduit considérablement son potentiel d'analyse et la possibilité de rejouer à l'identique la séquence. Cela provenait en partie de la difficulté technologique de réaliser l'interopérabilité des ressources, mais aussi d'une approche encore naïve de la pratique expérimentale en SIC. S'agissant pour l'essentiel d'observations de situations info-communicationnelles, il est étonnant que la fonction documentaire des métadonnées, dans ce cadre particulier, n'ait pas été mieux saisie à l'époque. Cela souligne également l'intérêt quasi exclusif porté au recueil des textes des échanges dont on estimait alors qu'il est porteur des informations essentielles à la compréhension des logiques d'interaction et des dispositifs d'information. On peut cependant atténuer cette critique en constatant que le potentiel créatif que représente la mise en œuvre expérimentale de situations d'interactions non contraintes n'a été que très peu envisagé à cette époque (Salber, Coutaz, 1993), (Le Bodic, 2005), (Ref.4). Elle a été davantage suivie dans le sens du *test de Turing*[23] pour évaluer la génération automatique de textes (réponses) (Meyer, Anis, 1992).

L'apport de cette expérience contractuelle est manifeste pour la suite de mes travaux. L'observation expérimentale m'a conduit à raisonner les problématiques de l'observation des pratiques info-communicationnelles effectives au travers de dispositifs spécifiques. Si l'observation demeure centrale, les objectifs lui étant associés diffèrent : il ne s'agit plus de créer les conditions d'une utilisation particulière mais de rendre compte des conditions courantes de l'usage. Ce changement de paradigme a comme conséquence la mise en œuvre d'une instrumentation qui n'a plus vocation à se substituer à un dispositif info-communicationnel mais plutôt à l'accompagner[24]. Il s'agit alors de capter et d'enregistrer tout ce qui est caractéristique de l'usage instrumental et de ses productions.

La dimension expérimentale demeure mais change de nature tant sur le cadre que les conditions de l'observation. Dans le contexte de la médiation numérique, le choix des observables disponibles n'est plus tout à fait le même, ainsi que les stratégies de captation qui caractérisent un échantillonnage. Ces choix ont des conséquences qu'il faut maîtriser en fonction des objectifs de la recherche. Ils se rapportent à des phénomènes plus ou moins ciblés et sont supportés par des représentations temporalisées produites par le dispositif d'observation, suivant une *granularité*[25] informationnelle plus ou moins fine. Les notions d'*observable* et de *représentation* numérique sont

[23] http://cogprints.org/499/

[24] J'utiliserai par la suite la métaphore de l'enveloppement (ou encapsulation) pour caractériser cet accompagnement.

[25] Caractérise le degré de description associé à la représentation. La finesse traduit une représentation très détaillée associée à des référentiels et des catégories nombreux. La granularité représentationnelle importante (fine) est susceptible de produire une représentation plus précise si les capteurs sont adaptés.

étroitement liées dans le contexte des dispositifs info-communicationnels. Déterminer un observable équivaut pratiquement à définir une structure de données qui s'instanciera dans des représentations au cours de l'observation. Il s'agit de « construire » une *collection*[26] de données qui soient fiables (donc vérifiées) et exploitables en fonction d'un contexte théorique.

Dans le contexte expérimental qui est le nôtre, l'activité expérimentale consiste tout d'abord dans une boucle rétroactive tout au long du processus d'instrumentation pour définir et mettre au point le dispositif d'observation en l'inscrivant dans un protocole d'observation. D'une manière plus classique, elle intervient ensuite comme moyen d'évaluer des hypothèses associées au projet scientifique portant l'observation.

Ces deux facettes de l'expérimentation sont indissociables car les données de l'observation ne peuvent pas s'interpréter en dehors du contexte de leur production. Le fait d'être impliqué dans la réalisation des dispositifs en amont de leur mobilisation m'apporte une connaissance, notamment des limites associées aux données issues de l'observation.

Ce mémoire d'HDR (deuxième partie) est l'occasion pour moi de souligner l'importance méthodologique de cette articulation dans une démarche de construction empirique des connaissances.

2. Une trajectoire en Sciences de l'information et de la communication

2.1 Une inscription disciplinaire.

J'ai occupé un poste d'ATER (1993) avant d'être recruté (1994) sur un poste de Maître de Conférences en Sciences de l'information et de la communication (71e section) au sein du département information-communication de l'IUT2 de l'Université Pierre Mendès France – Grenoble 2 (UPMF).

Dès lors, je me suis investi dans la mise en œuvre d'enseignements théoriques et pratiques liés aux Technologies, à leurs développements et leurs incidences sur les mondes professionnels (au sens d'Howard Becker) et sur la société en général. Mes enseignements ont concerné prioritairement des filières professionnelles de la communication d'entreprise (6 ans) et des métiers de l'édition et du livre (14 ans).

La mutation numérique engagée dans les professions de l'information-communication dans le contexte du développement rapide des technologies numériques de l'Information et de la Communication (TIC[27]), l'ouverture à la concurrence du marché des télécommunications[28],

[26] Une collection désigne un ensemble homogène, cohérent et pertinent d'entités. Nous adoptons ce terme en référence aux considérations documentaires qui sont les nôtres.

[27] Dans toute la suite, les Technologies évoquées seront de nature numérique. Dans l'évocation liée à la pédagogie, le terme TIC fait écho à l'usage familier de l'expression pour désigner des technologies émergentes, pas seulement numérique et malencontreusement souvent qualifiées de "nouvelles".

[28] Deuxième livre vert octobre 1994, janvier 1995

l'avènement de l'Internet grand public, m'a conduit à prendre en charge rapidement (1995) la direction du département information-communication pendant 5 ans. Ayant conduit deux mandats, j'ai activement contribué à la définition des compétences TIC dans l'évolution des programmes pédagogiques nationaux. Suivant cette dynamique d'implication dans les filières de l'information-communication, j'ai lancé la création d'un Institut Universitaire Professionnalisé (IUP) *Métiers des Arts et de la Culture – métiers du livre* (co-habilité en 1998) qui a ouvert ses portes conjointement à Nanterre (Paris X) et à Grenoble (UPMF) à la rentrée 1999. J'ai dirigé cet IUP durant les 8 années avant d'envisager son intégration dans les logiques LMD au sein de l'Unité de Formation et de Recherche (UFR) Sciences Humaines et Sociales (SHS) de l'université Pierre Mendès France en septembre 2007. Par la suite, j'ai assuré durant 6 ans, la responsabilité des spécialités professionnelles de Masters de l'UFR SHS qui en sont aujourd'hui encore les déclinaisons[29].

L'exercice des responsabilités et mon parcours d'enseignement m'ont donné une vision très transversale et une lecture stratégique des enjeux de formation liés à l'information-communication et aux métiers de ce secteur. Cela m'a incité à développer des partenariats de type recherche-action avec des entreprises ou des organismes de représentation professionnelle et à être au plus près du terrain. Les raisons stratégiques et politiques n'en sont pas les seuls arguments. J'ai toujours considéré cette immersion dans les mondes sociaux et leurs problématiques comme indispensable pour nourrir le métier d'enseignant-chercheur.

La dissolution du CRISS, concomitante avec la fin de ma thèse, a localement conduit à une recomposition du panorama de la recherche en Sciences de l'information. Le débat scientifique qui s'est fait jour portait sur la place de la langue naturelle envisagée comme cadre directeur (mettant ici l'accent sur l'analyse automatique du document) ou bien comme outil mettant, cette fois, l'accent sur l'interrogation (*i.e.*: le système de recherche). Dans le premier cas, la compréhension des mécanismes linguistiques prime alors que dans le second c'est l'efficacité. Ayant une approche guidée par les logiques et la pragmatique de l'interaction homme-ordinateur multimodale, j'ai privilégié cette seconde voie.

Durant cette période, mes travaux se sont orientés du côté des Sciences de l'Information. En effet, les premiers services documentaires mis en ligne sur Internet (wais, gopher, etc.) introduisaient une conception très différente des systèmes d'information documentaire (Ref.6-7). La transformation s'annonçait radicale, imposant de penser la notion de médiation documentaire et son instrumentation dans un cadre nouveau : le Web.

La mise en place sur le bassin grenoblois d'un réseau documentaire de coopération REDOC (1994-2004) va constituer un cadre d'observation privilégié de ces changements (Ref.8). À partir de 2000, j'ai souhaité développer un projet scientifique mieux articulé aux domaines et problématiques de formation de l'IUP (Ref.9-15). L'axe scientifique que j'ai privilégié portant sur les pratiques info-communicationnelles émergentes associées aux développements des ressources du Web et des

[29] Master professionnel : (2011) *Production et Médiations de l'œuvre,* (2013) *Productions et Médiations des Formes Culturelles*. L'accréditation a été obtenue pour l'ouverture en 2016 du Master de site : *Métiers du livre et de l'édition.*

modalités de son accès, j'ai alors rejoint le CEDPIC (Centre d'Étude des Dispositifs et des Processus d'Information-Communication), jeune équipe dont le projet a porté sur l'analyse de la communication à l'ère des réseaux. L'avènement de la 3G et des smartphones marquait cette époque et a orienté ce projet scientifique. Les débits et les capacités des smartphones ont introduit une qualité propice à la diffusion multimédia des contenus et aux pratiques info-communicationnelles nomades individualisées. Les problématiques étaient alors centrées sur l'évolution des modalités temporelles et spatiales de consommation de ressources et d'accès aux services info-communicationnels.

Étudier ces pratiques suppose une capacité d'observation permanente y compris à l'occasion de mobilités et donc en dehors de l'espace maîtrisé du laboratoire. Pour financer ce programme de recherche, j'ai répondu et obtenu un contrat de financement de l'ACI Ville (Ref.13). L'objectif de ce financement était de réaliser une plateforme expérimentale, baptisée PLEXUS[30] (cf. *infra*), permettant l'observation écologique et continue des usages individuels des navigateurs Web embarqués sur les smartphones, et dans différents types de dispositifs mobiles. À la suite de son développement, PLEXUS a été mobilisé dans un programme de recherche-action mené avec l'UMR PACTE s'appuyant sur un contrat industriel nous liant (CEDPIC, PACTE) à France Télécom Recherche et Développement (Ref.16-17). Ce travail a jeté les bases d'une collaboration plus poussée avec le laboratoire PACTE et plus précisément avec les équipes de l'axe "territoire".

Les logiques de l'organisation de la recherche appelant un regroupement des forces dans des unités plus importantes, l'équipe du CEDPIC a été appelée à intégrer l'UMR-PACTE en 2007. Le projet d'équipe que nous avions porté a été l'occasion d'incuber un programme de recherche autour des dispositifs et des processus d'information-communication et de l'implanter en tant que nouvelle thématique au sein du laboratoire PACTE.

Depuis cette date, je contribue activement au développement de ce programme. Il articule des projets interdisciplinaires impliquant suivant leur nature, des géographes, des politologues et des sociologues.

2.2. Une démarche exploratoire et inductive

L'ordinateur individuel qui a marqué un tournant dans l'évolution des Technologies de l'Information Communication au début des années 1980 est maintenant remplacé dans ce rôle par des configurations variables d'objets connectés. Le poste de travail qui caractérisait la situation privilégiée d'utilisation de ressources informatiques est devenu un environnement permanent et personnel. En miroir de cette évolution technologique, l'usage social des technologies numériques s'est largement diffusé. Il affecte désormais l'ensemble des activités professionnelles ou domestiques, privées ou publiques, individuelles ou collectives.

[30] Dont l'implémentation informatique a été réalisée grâce aux efforts d'Eric Guinet que j'ai pu salarier en tant qu'ingénieur d'étude sur ce contrat.

L'instrumentation des facettes informationnelles et communicationnelles de l'activité humaine n'est pas une nouveauté. La convergence numérique et l'Internet en particulier prolongent une tendance qui est portée par notre société dans son rapport à la technique soulignée dès l'après-guerre.

Ainsi, l'instrumentation s'impose de plus en plus comme moyen individuel d'engagement dans l'activité collective et donc sociale. Elle se justifie dans une rationalité opérative et constitue dans le même temps une injonction à laquelle il est difficile d'échapper. Cette injonction est double. Elle est pour partie celle d'un public, d'une audience qui se constitue dans la pratique de l'outil établissant au sein de ce collectif les normes circonstancielles d'usage. Elle est également le fait d'un cadre socio-technique sous-jacent qui assure la consistance du dispositif.

À partir des années 2000, l'importance du Web dans les pratiques info-communicationnelles va m'inciter à développer une approche instrumentale de l'observation située suivant des préoccupations distinctes et complémentaires : l'analyse écologique des pratiques du Web (§ 2.2.1), l'exploration des flux de données individualisés (§ 2.2.2), ainsi que la marchandisation et l'industrialisation des traces personnelles (§ 2.2.3)

Le développement chronologique de ces trois points souligne l'évolution de travaux et de recherche-actions qui s'inscrivent dans un contexte scientifique et sociétal de plus en plus sensible à la question des données et des *traces d'usage*. Cette évolution s'est concrétisée dans la réalisation de dispositifs d'observations qui sont moins liés au poste de travail individuel et davantage en interface avec les services du Web à partir desquels l'activité collective s'organise. Ainsi progressivement, la problématique de l'individu s'est dissoute dans une problématique d'identité et d'interaction au sein d'un groupe.

2.2.1. L'analyse écologique des pratiques du Web

Au début des années 2000, les recherches en SHS portant sur les usages sociaux du Web ont alimenté une pensée sociologique qui s'est constituée dans les années quatre-vingt autour de la notion d'individualisme[31] (Dumont, 1983). Plusieurs lectures relatives au Web et à l'Internet coexistent alors (Casilli, 2010). Historiquement, la première approche est celle d'une désocialisation (Wolton, 1997). Mais progressivement, une autre lecture s'est fait jour, que l'on peut résumer dans le titre de l'article de Patrice Flichy : L'*individualisme connecté* (Flichy, 2004). Celui-ci traduit le paradoxe d'un isolement de plus en plus grand dans la conduite de l'activité ordinaire dans les sphères privées ou professionnelles, alors que dans le même temps les logiques de réseaux se développent (Jouët, 1997). Cette rupture de paradigme dans la compréhension des usages sociaux du Web justifiait l'analyse des pratiques effectives et individuelles du Web, en situation "naturelle" (écologique). C'est précisément l'objectif fixé à la plateforme PLEXUS conçue pour capter et traiter les comportements individuels en grand nombre.

[31] Ne porte pas sur le versant de l'individualisme méthodologique défendu par Raymond Boudon, mais plutôt sur les travaux en Sociologie de la famille (cf. travaux de François de Singly).

L'architecture de PLEXUS permet de connecter un grand nombre de clients[32] au serveur. Toutes les informations qui transitent par le navigateur sont ainsi capturées et renvoyées au serveur. Par ailleurs, toutes les actions de l'utilisateur sur le navigateur, liées à la consultation (défilement, etc.) ou à l'édition (touches claviers, etc.) de contenus, sont également renvoyées et journalisées sur le serveur[33].

PLEXUS a été mis en œuvre à l'occasion d'un projet questionnant le rapport entre *TIC[34] et développement durable des territoires de montagne* (Ref.16). Les machines publiques de deux espaces publics numériques (EPN) ont ainsi été équipées du client PLEXUS[35] pour une expérimentation qui dura 6 mois. Une fois la plateforme validée, les difficultés rencontrées ont été moins d'ordre technique que d'ordre méthodologique. En effet, une partie de l'enregistrement concernait des pages html dans une logique et des difficultés équivalentes à celles de l'archivage du Web. Une session internaute comportait un très grand volume d'informations semi-structurées associant des données de navigation (pages) et des données de navigateurs (micro-événements). La fouille de ces données a constitué un problème technique important. Il a nécessité le développement "à façon" d'extractions compatibles avec des formats, des méthodes et les limites des outils d'analyse classique (statistique, statistique textuelle). Plus globalement, l'alignement réflexif entre questionnements scientifiques et jeux de données recueillis a soulevé des enjeux méthodologiques et disciplinaires difficiles à dépasser dans le cadre d'un seul projet. L'instrumentation a ses propres logiques de développement et ouvre des potentialités de capture/mesure dépassant le cadre négocié du projet d'expérimentation. L'apport de la mise en pratique souligne ces écarts et permet des reformulations de questionnement scientifique *ex post* appelant d'autres expérimentations. C'est alors dans la dimension diachronique de l'expérimentation que se construit aussi le bénéfice scientifique.

Dans le cas présent, PLEXUS s'est inscrit dans un dispositif complexe d'analyse l'associant à d'autres méthodes plus classiques, telles que l'observation participante, l'entretien et l'enquête. Articuler les différents protocoles et les données récoltées est l'un des apprentissages de cette expérimentation.

---

[32] Équipements individuels multiples pour un grand nombre d'individus.

[33] Nous avons essentiellement travaillé dans l'environnement windows et avec IE qui présentaient l'avantage d'être déployés sur les ordinateurs et certains smartphones.

[34] Le terme désigne ici les infrastructures technologiques majoritairement numériques assurant l'accessibilité aux ressources info-communicationnelles (dont l'Internet) dans des conditions spatiales (topographiques) défavorables aux techniques déployées par les opérateurs pour couvrir le territoire. Ces infrastructures conditionnaient des aménagements territoriaux spécifiques amenant une politique d'investissement et de services adaptée aux modes de vie et de "consommation" du territoire. En particulier dans la définition de points d'accès publics (cyber-centres, Espaces Publics Numériques, etc.). De ces aménagements, découlaient des modalités d'usage et de pratiques que nous souhaitions étudier.

Les apports spécifiques de PLEXUS au questionnement "territorial" du projet, portent principalement sur la nature des ressources documentaires mobilisées et la nature des activités qui leurs sont associées au cours des sessions.

[35] Les internautes étaient avertis par un pop-up de l'expérimentation en cours et pouvaient refuser le suivi de leur session qui sinon se déclenchait (entre 35 % et 50 % d'acceptation suivant les centres).

L'historisation des micro-événements du navigateur a démontré son intérêt dans la reconstruction de la ligne temporelle des séquences d'activité au cours des sessions[36]. L'*accountability* de la trace reconstituée est bien vérifiée, sa lecture est cohérente et permet l'analyse "par-dessus l'épaule" de l'internaute. Cette représentation de session utilisateur permet d'aborder l'analyse des comportements suivant les trois niveaux (*information searching, information seeking behavior, information behavior*) décrits par Thomas Daniel Wilson (Wilson, 2000).

2.2.2. L'exploration des flux de données individualisés

Après 2004, l'arrivée des plateformes info-communicationnelles désignées comme *Réseaux sociaux numériques* (RSN) marque une autre évolution.

À l'instar de Facebook, considéré comme l'un des pionniers, ces plateformes canalisent et structurent un ensemble de services tiers permettant : de définir un profil personnel, de gérer des relations et de naviguer dans des contenus publiés suivant différents niveaux de publicisation (Ellison, Boyd, 2007), (Stenger, Coutant, 2010). Dans ce cadre, l'interaction individuelle avec la plateforme est l'ingrédient d'une interaction relationnelle et éditoriale entre pairs. En attribuant une forte valeur, notamment symbolique, à l'inscription dans des espaces de relations socialisées, ces plateformes fidélisent leurs abonnés et en dynamisent l'activité.

Pour ces plateformes, la captation de la valeur réside moins dans la production collaborative que dans l'enrichissement des représentations individualisées basées sur la durée de la relation et l'étendue des interactions individuelles. Les traces d'interaction ont acquis le statut de données et font l'objet d'une journalisation. Elles portent les espoirs d'une caractérisation comportementale individuelle très fine, propre à répondre aux-enjeux d'une économie d'individualisation de masse (Da Silveira & al., 2001), (Ref.18/19).

L'émergence des réseaux sociaux numériques va provoquer un vif intérêt dans les Sciences sociales et va entraîner la mise en place, au sein du laboratoire, d'un groupe de travail : *Mediacorpus*. Ce groupe a été l'occasion de croiser les points de vue disciplinaires sur les données numériques et sur l'analyse des phénomènes médiatiques. C'est à partir de cette réflexion commune sur l'approche empirique que mon programme de recherche va s'ouvrir aux thématiques de l'espace public et des phénomènes médiatiques, et que des collaborations durables avec des sociologues et des politistes verront le jour (Ref.21-22, 24-28, 30-35, 37-41).

L'analyse en masse des comportements info-communicationnels accompagnant les événements médiatiques, suppose de développer des représentations et des traitements constituant des formalisations individualisées. Ces développements peuvent être pertinents dans l'élaboration d'une offre de services.

Or la plateforme PLEXUS ne répond pas pleinement à ces enjeux. En conséquence, j'ai développé une plateforme spécifique : MEDIASWELL qui permet d'organiser un archivage simultané de différentes sources de données, interrogées suivant des scripts programmables. MEDIASWELL permet de fonctionner suivant plusieurs logiques : l'établissement d'une communication avec les

[36] La réception asynchrone des composants multimédia et des pages html ne permet pas cette reconstruction fidèle.

plateformes fournissant les données au travers d'API (par exemple Twitter); l'extraction (*Scraping*) permettant l'analyse des pages de sites ressources à partir de leur spécification ; la syndication de flux de type RSS. MEDIASWELL est également couplé avec des outils de traitements automatiques de la langue (TAL) développés en collaboration avec le Laboratoire de Linguistique et de didactique des langues étrangères et Maternelles (LIDILEM).

MEDIASWELL a été mis en œuvre à l'occasion de nombreux événements médiatiques tels que : les jeux Olympiques (Hiver 2010, Eté 2012, Hiver 2014) (Ref.22, 32, 33), les élections présidentielles de 2012 (Ref.21, 24, 25, 27, 28, 31, 33, 35,37, 38), les élections européennes de 2014 (Ref.39). Aujourd'hui, ce dispositif d'observation et d'analyse est impliqué dans le cadre de l'ANR RSJ-MéDiS[37] à laquelle je participe.

La mise en œuvre de cette plateforme dans des contextes récurrents a permis de faire évoluer la réflexion méthodologique. En particulier, la constitution d'un contexte documentaire associé à la collection a donné lieu à une collaboration suivie avec le département du dépôt numérique de la Bibliothèque Nationale de France (BNF) en charge de l'archivage du Web.

La plateforme MEDIASWELL intègre les éléments d'une réflexion qui s'oriente vers l'exploration instantanée et simultanée des données issues des services du Web (Ref.21-39). Il s'agit toujours de constituer une mémoire durable à des fins exploratoires des pratiques individuelles, associées désormais à un environnement multi-services.

Cependant, l'objectif comporte une autre dimension : il s'agit aussi, de constituer simultanément une mémoire de travail instantanée à partir de laquelle peuvent s'envisager et s'évaluer des outils et des méthodes en vue d'une analyse immédiate et longitudinale de l'événement. D'ailleurs, cette problématique du *juste à temps*[38] est aussi celle des industries d'exploitation analytique des données du Web (*Web analytics*) auxquelles les professionnels de l'opinion (instituts de sondages, médias) se réfèrent de plus en plus (Ref.25, 31). Concomitamment, l'importance stratégique croissante que l'industrie accorde aux données du Web conduit les entreprises de services numériques détentrices de ces données (Twitter, Facebook, LinkedIn, etc.) à restreindre fortement les services d'accès public à leurs données (API[39]) afin d'en préserver la valeur.

C'est cette tension sur la disponibilité des données de services qui m'a conduit à m'intéresser aux problématiques de l'*espace de publication* des données du Web. L'objectif est d'étudier la porosité des espaces de données de ces entreprises et les modes de diffusion de l'information dans les différents espaces privés et publics qu'ils innervent. Par ailleurs, il s'agit de contourner l'effet "boite noire" que le Web engendre c'est-à-dire de raisonner l'ouverture des méthodes et des données du Web, dans un contexte d'espace public comme un enjeu pour les sociétés démocratiques contemporaines.

---

[37] ANR 15-CE26-0006-01 : *Responsabilité Sociale des Journalistes – Médias Diversité Sport*. Voir le document annexé : *Curriculum Vitae*.

[38] Voir à ce sujet le cycle de conférence mis en œuvre en 2012 à l'UNIL et à l'EPFL (Lausanne) dans lequel s'inscrivait la présentation de nos travaux sur l'archivage de Twitter. http://wp.unil.ch/digitalera/rencontres-2012/

[39] *Application Program Interface* - Interface d'accès aux données d'application.

Sur le plan méthodologique, la mise en œuvre de dispositifs expérimentaux dans un contexte empirique m'amène à m'interroger sur les conditions d'analyse d'informations fragmentées et d'une granularité très fine. Plusieurs axes de travaux en découlent :

1) Le fait de considérer le Web comme un espace de données représentatives d'individus et de groupes à une très large échelle présente d'incontestables difficultés du point de vue des méthodes d'analyse couramment utilisées dans les Sciences sociales et humaines. La disponibilité de données en nature et en quantité jusqu'alors inaccessibles est tout autant un problème que celui de leur qualité ou de leur représentativité. Dès lors, il est nécessaire pour accompagner la démarche empirique de fonder ou refonder des méthodes d'investigation et d'analyse et de proposer des outils adaptés. Cela suppose de questionner en permanence les méthodes employées et d'en vérifier l'adéquation et les limites ;
2) Le plus souvent, les données brutes[40] (*raw data*) issues de la collecte ne constituent pas celles de l'analyse. Elles ne sont qu'un intermédiaire, sous-déterminé par rapport à leur exploitation scientifique dont elles doivent cependant anticiper les contraintes. L'indétermination du statut des données est renforcée par le décalage entre la mise en œuvre de l'expérience et l'étude à proprement parler, ainsi que par les difficultés à rejouer ou à maintenir durablement une situation expérimentale. Cette indétermination tient à la nature de la démarche empirique et exploratoire. La dimension exploratoire, caractérisée par le fait que l'observation se déroule en même temps que la construction de l'objet d'étude, est une caractéristique de mon activité de recherche.
3) Ces considérations sur les données brutes m'ont amené à m'intéresser aux problématiques de la constitution de collections à finalités scientifiques en dehors d'un cadre de recherche précis. Elle pose la question de la mise en cohérence des collections expérimentales avec les collections institutionnalisées du Web, comme celles que la Bibliothèque Nationale de France (BNF) ou l'Institut National de l'Audiovisuel (INA) ont pour mission de constituer. Les questions sous-jacentes, concernent l'autonomie relative des données brutes et les logiques de standardisation et de communication scientifique.

Le contexte d'événements médiatiques associés à des temps forts de la vie sociale m'a fait prendre conscience des interactions fortes entre l'observation d'une part et le projet scientifique d'autre part. L'évidence est moins flagrante qu'il n'y parait. En effet, dans de nombreux cas, l'observation est souvent une manière d'aborder des situations d'usage nouvelles (ce fut le cas pour nous) pour lesquelles nous sommes désarmés. La plus part du temps, la formulation des hypothèses est succincte et les observables peu définis. C'est dans la démarche explorante et l'itération expérimentale que les hypothèses vont être formulées et les variables définies, raccrochant ainsi l'observation aux projets analytiques. Cependant, la conduite d'analyses souvent exploratoires dans le contexte empirique ne construit pas un projet scientifique, du moins au départ.

[40] Le terme de *données brutes* est à prendre au sens de données "immédiates" obtenues (plutôt que produites) sans garantie de validité et de pertinence par rapport à l'analyse envisagée. Je reviendrai sur ces caractéristiques dans la suite du mémoire.

Dans mon cas cette ligne directrice est contenue dans une pensée critique inspirée des travaux de Jürgen Habermas.

2.2.3. Marchandisation et Industrialisation des traces personnelles

La proximité problématique avec les industries du Web des données est une opportunité que j'ai saisie en participant à l'ERT UmanLab et en devenant conseiller scientifique pendant 5 ans auprès de la start-up du Web PRÉDICTYS[41] (2005-2010). Le projet d'entreprise portait sur l'enrichissement de profils individuels et le ciblage fin d'*emails*. Les conditions d'activité de cette société, au statut d'agence Web incorporant également des compétences de routage, sont typiquement celles du *Web Usage Mining*, point de rencontre du *Data Mining* et du *Big Data* sur le Web.

Dans les dernières années, la partie sur laquelle j'ai travaillé concernait la définition d'une composante bien précise du système d'information (SI), à savoir le PCA (*Producer Consumer Analytics*). Cette composante est dédiée à des fonctions analytiques se rapportant aux flux d'information de production et de consommation. Ici, l'un des objectifs innovant du PCA visait la mise en œuvre d'une analyse comportementale. Dans le contexte de l'e-mailing publicitaire, l'analyse comportementale a pour objectif, en premier lieu, de maintenir élevé le retour sur investissement d'une campagne publicitaire tout en limitant le nombre de personnes contactées et les risques associés au routage. Cette logique qualitative traduit une moralisation/régulation du marché de l'e-mailing publicitaire, limité par les pratiques massives et les contrôles de flux des fournisseurs d'accès. En second lieu, l'intérêt de ce ciblage est de maintenir durablement une relation avec le consommateur (destinataire), condition nécessaire pour affiner le profilage individuel et enrichir la caractérisation personnelle.

L'intérêt scientifique de cet investissement portait sur la mise en relation du cycle de vie de la relation internaute (consommateur) avec le cycle informationnel ainsi que sur la nature prédictive de la caractérisation comportementale. Ces travaux se concrétisent aujourd'hui dans la généralisation des pratiques de la gestion de relation client (CRM) dans les entreprises.

Cette problématique n'est pas éloignée de celles de l'interaction dialogique. Les e-mails publicitaires concernés sont assimilables à des pages Web, contenant des textes courts et des liens, indexées par un jeu de métadonnées situant le texte dans un espace d'une dizaine de dimensions se rapportant à des thématiques, à des valeurs et d'autres catégories. La nature des contenus est variable, il peut s'agir, par exemple, de messages publicitaires, des jeux concours, des promotions et des avantages clients. La caractérisation comportementale doit avoir une fonction prédictive permettant de constituer la cible ayant le meilleur potentiel dans les termes du retour sur investissement (ROI) ou de la caractérisation elle-même. Elle doit permettre de prévenir des attitudes de rejet (black-listage) dont les conséquences économiques sont critiques.

La modélisation comportementale devait dans le cas présent se plier aux exigences du *Big Data*. Il s'agissait en effet, d'actualiser la caractérisation de quelques 500 000 individus ciblés journellement

[41] http://www.predictys.fr

pour le compte de campagnes différentes adressant des bases de plusieurs millions d'internautes pour chacune d'elles et cela en moins de vingt-quatre heures.

L'approche comportementale envisagée est *endogène*, c'est-à-dire qu'elle ne repose, à chaque instant et pour une campagne publicitaire, que sur l'ensemble des historiques individuels de couples, "e-mail – séquence d'actions de réception". La séquence d'actions est une chaine temporelle construite à partir des données de *logs* (traces) associées aux actions élémentaires de l'utilisateur sur tout *email* envoyé. Ces actions sont asynchrones par rapport à l'envoi ou la réception de l'*email*. Un *email* reçu peut être ouvert plusieurs fois, à différentes dates et horaires, sur une période indéterminée. Il en va de même pour les liens contenus dans l'*email*, qui peuvent être parcourus ou non.

Les limites de l'approche endogène m'ont amené à m'intéresser à l'exploration contextuelle des données *exogènes*, c'est-à-dire aux données du Web qui sont liées aux données endogènes. Cette perspective m'a incité en particulier à renforcer les collaborations engagées avec la BNF sur les problématiques de l'archivage institutionnel en développant l'exploitation des archives constituées (Ref.22, 26, 29, 34, 36, 40).

L'intérêt de la collaboration avec PRÉDICTYS (5 ans) a été sans conteste la possibilité de travailler dans un cadre industriel, en situation de production au cœur même de l'innovation socio-technique et de l'économie des données massives (Ref.23). Il a été également celui d'une réflexion sur la chaine de la valeur (Ref.18, 19) et sur l'éthique de l'analyse des données personnelles (Ref.20).

3. Un positionnement scientifique

Dans les deux parties précédentes, j'ai voulu traduire la dynamique de mon parcours d'enseignant-chercheur en Sciences de l'information et de la communication afin de souligner comment les objets de ma recherche se sont construits. Je me propose dans cette troisième partie, de mettre en perspective les éléments scientifiques qui caractérisent mes recherches et à partir desquels se développe ce mémoire.

Les Sciences de l'information et de la communication se sont fondées et revendiquent toujours l'interdisciplinarité des approches dans la constitution du champ d'étude qu'elles déterminent. Pour Robert Boure, le «*noyau dur [de la discipline] est constitué par l'étude des médias et plus généralement des techniques, des dispositifs et des acteurs de l'information et de la communication*» (Boure, 2002 p22). Mes travaux se situent sans ambiguïté dans le sous ensemble numérique des TIC.

Le concept de *dispositif* qu'utilise R. Boure, et que j'ai employé jusque-là suivant le sens commun, mérite une attention toute particulière. Ce terme restitue les interactions complexes entre : les différents éléments technologiques mobilisés configurant une *machine à communiquer*[42], les acteurs de l'espace social qui l'opèrent et enfin les protagonistes impliqués dans son utilisation. L'adoption du terme met l'accent sur une dimension temporelle beaucoup plus soutenue. En effet, un dispositif peut avoir une grande stabilité comme au contraire n'avoir qu'une existence éphémère. Il est

[42] Terme proposé par Pierre Schaeffer à une époque où les objets techniques *concrétisaient* une unité fonctionnelle au sens de Gilbert Simondon (Schaeffer, 1970), (Simondon, 1958).

tributaire de configurations sociales et techniques activées ou activables dans la situation de sa mobilisation.

La contingence soulignée par ce terme (*dispositif*) me parait devoir être intégrée dans l'analyse des usages des TIC. Il détermine un niveau d'appréhension des situations qui reste en retrait par rapport aux spécifications des objets communicants qui le constitue. Ce décalage s'impose, à mon sens, pour plusieurs raisons de nature méthodologique. Cette notion introduit une articulation qui devient nécessaire compte tenu de l'incorporation toujours plus grande des technologies communicantes dans l'environnement quotidien. Par ailleurs, la réification apparente que constituent les classes d'objets (mobiles, smartphones, etc.) ou de services (ceux du Web en particulier) incite à des généralisations hâtives dont il faut se prémunir. Le terme même de TIC participe de cette illusion d'unité, essentiellement fonctionnelle, dont il faut se méfier. À la différence, le terme de dispositif invite à une mise à distance en engageant une compréhension socio-technique de sa réalisation[43]. En décalant le discours, le terme *dispositif* atténue le risque d'un déterminisme trop souvent associé aux technologies.

L'usage de ce terme n'est cependant pas sans écueils et a déjà fait l'objet de nombreuses critiques et tentatives d'appropriation, notamment dans le champ des SIC. Je reviendrais plus longuement sur cette question dans le second chapitre de ce mémoire, en cartographiant notamment son usage dans notre discipline. Donner un sens plus social à ce terme permet de développer également une pensée critique. Dans ce sens, je m'inscris dans la continuité des travaux de M. Foucault se rapportant aux enjeux de pouvoir qui interviennent dans la concrétisation des dispositifs info-communicationnels.

L'émergence relativement récente de cette notion dans notre discipline me parait devoir être approfondie et encouragée. J'y contribue dans ce mémoire en la justifiant d'un point de vue méthodologique. Au travers de ce concept, c'est une démarche empirique qui est mise en avant dans une recherche qui interroge de manière privilégiée les médiations que les dispositifs info-communicationnels supportent.

Par ailleurs, l'étude des médiations info-communicationnelles doit s'envisager dans les contextes sociaux de l'activité globale qui la justifie. Il y a là une nécessaire ouverture interdisciplinaire dans la mesure où la démarche d'analyse porte, de fait, sur des faits sociaux. C'est dans ce sens que j'ai participé à des projets tels que, par exemple, TriElec[44] à l'occasion de la dernière campagne électorale de 2012. Les travaux que j'ai produits sur Twitter en collaboration avec Françoise Papa s'inscrivent dans ce contexte.

L'approche que j'adopte repose sur l'hypothèse que les dispositifs info-communicationnels au travers de leurs médiations ont un rôle dans la constitution de ces faits sociaux et qu'ils ne sont pas neutres. Ils contribuent à une structuration de l'espace social au travers des logiques d'usage qui

[43] Voir ci-après § 3.2.

[44] https://sites.google.com/a/iepg.fr/trielec/resultats-analyses/enquetes-pre-electorales

s'instaurent. L'effet de structuration introduit par l'usage de Twitter durant la campagne électorale de 2012 est à ce titre remarquable (Ref.25, 27, 28).

L'absence de neutralité du médium, que soulignait déjà Marshall McLuhan dans les années 1960[45], est un élément qui n'apparait pas de manière toujours évidente dans les travaux. La métaphore du "tuyau" est latente dans bon nombre d'études de phénomènes sociaux, négligeant ainsi les effets des médiations techniques qui les traversent. Cette absence de neutralité est cependant le point de rencontre d'une recherche qui s'organise autour de la question des usages[46] et dont Josiane Jouët donne une excellente synthèse (Jouët, 1997).

Pour moi, l'*influence* est d'abord celle qui s'exprime, tenant compte d'arbitrages individuels, dans l'instrumentation de l'activité. L'attribution d'une capacité d'arbitrage revient à reconnaître à l'individu une autonomie au moins suffisante pour qu'en dehors des cadres d'obligation explicite (i.e situation de travail par exemple[47]), la décision d'*instrumentation* et les modalités de celle-ci soient le reflet de *choix rationnels*[48].

Elle est aussi l'*influence* que l'on peut associer aux agents qui opèrent au travers des dispositifs, des enjeux de pouvoir sociaux, économiques et politiques eux-mêmes institutionnalisés et reflétant des intérêts partisans ou l'intérêt commun au travers des référentiels normatifs (lois, normes sociales, techniques, etc.). Ce jeu d'influences dynamiques s'exprime au sein du dispositif et participe à l'équilibre interne des tensions que les mises en relation entre acteurs établissent.

Elle est enfin l'*influence* d'un dessein que poursuivent les acteurs sociaux utilisant les potentialités qu'offrent ces dispositifs comme moyen d'exprimer et de diffuser leurs opinions ou idéologies dans la perspective de mobiliser une audience au sein de l'espace public. Mon approche de l'*influence* est inspirée des travaux de Elihu Katz et Paul Lazarsfeld (Katz, Lazarsfeld, 2008).

Ce sont en définitive deux conceptions, sociale et cognitive, du *sujet* qui sont mobilisées dans l'approche liée aux pratiques effectives que je poursuis. Cette double analyse est nécessaire car de plus en plus, la médiation de l'activité se double d'interactions interpersonnelles dans un contexte collaboratif - allant même jusqu'à en constituer le cœur d'activité.

Ainsi, la compréhension des faits sociaux implique celle des dispositifs et de leur mobilisation dans l'établissement de ces faits. Cependant, cela ne signifie pas que la compréhension, même contextualisée, qui peut être associée à la mobilisation et l'usage du dispositif soit suffisante pour saisir les faits sociaux dans toute leur étendue et leur complexité. En revanche, ce glissement permet un positionnement disciplinaire clair et une orientation cohérente d'un travail qui se concentre sur les pratiques info-communicationnelles *médiatées* ou *médiées*[49].

---

[45] «…*c'est le médium qui façonne le mode et détermine l'échelle de l'activité et des relations des hommes*.» (McLuhan 1968, p24).

Replacer dans la dynamique de l'info-com. Comment appréhender l'influence, etc. dans une histoire disciplinaire.

[46] La question des usages est abondamment traitée dans ce mémoire (chapitre 4).

[47] Situations que je n'ai pas abordées dans mes travaux.

[48] Cette notion sera abordée dans un chapitre consacré aux logiques de l'usage (chapitre 4).

[49] Sens développé par Rachel Panckhurst (Panckhurst 2006, 2007)

Par la suite, la réflexion que je poursuis dans ce mémoire, repose sur trois axes majeurs que je développe ci-après.

3.1. Le Web comme espace de pratiques

La densification et la diversification des pratiques info-communicationnelles ont fait émerger des enjeux économiques, sociaux, politiques et culturels majeurs. Le World Wide Web (Web) et ses développements cristallisent ces enjeux. C'est la raison pour laquelle je me suis plus particulièrement concentré sur les dispositifs reposant sur les technologies et les services du Web. La nature intégratrice du Web, la standardisation poussée qu'il introduit et son usage largement répandu sont à l'origine de ce choix. Force est de constater d'ailleurs que certaines pratiques associées au Web constituent par elles-mêmes des faits sociaux.

La définition technique du Web qui a prévalu et qui le décrit comme un système hypertextuel fonctionnant sur Internet n'est pas satisfaisante. Les catégories éditoriales proposées telles que site, blog, forum et plus récemment réseaux sociaux ne donnent qu'un aperçu partiel des enjeux info-communicationnels qui lui sont associés.

Le Web ne peut définitivement plus être considéré comme un espace homogène. Il est caractéristique de ce que Jean-Louis Le Moigne dénomme un *système complexe* (Le Moigne, 1977). On pourrait me semble-t-il, pour être plus précis, reprendre le terme d'*hyper-complexe* ou dit autrement, un *complexe de systèmes complexes,* expressions qu'utilise Edgar Morin pour décrire le cerveau (Morin, 1986, p97). Toujours en suivant E. Morin, on peut encore caractériser le Web de *système ouvert* (Morin, 1990, p29). Cette perspective, issue de la thermodynamique, qu'E. Morin adopte pour décrire le vivant, me parait féconde. L'approche qu'il propose intègre une relation d'interdépendance dynamique du système et de son environnement c'est-à-dire dans le cas présent, la société globale. C'est un lien qu'il qualifie d'« *absolument crucial »* sur l'ensemble des plans épistémologique, méthodologique, théorique et empirique (Morin, 1990, p32).

Par ailleurs, au cours de ces vingt années, la marchandisation progressive et la production industrialisée des données du Web ont introduit une fragmentation en sous-espaces plus ou moins perméables les uns avec les autres. Cette fragmentation a une incidence significative et induit différentes lectures du Web. L'une d'entre elles oppose un espace public (numérique) au sens d'Habermas qui s'inscrit dans la continuité idéalisée, d'un Internet grand public, aux espaces privatisés regroupant des *bulles*[50] propriétaires administrées par des plateformes de services (Google, Amazon, Facebook, etc.), (Mattelart, Vitalis, 2014). La proximité du Web avec la notion d'espace public est renforcée d'autant plus que la présence élargie des médias d'information sur le Web – constitutive d'un espace médiatique (numérique) – le connecte à la notion d'espace médiatique, voire également à celle de démocratie (électronique).

Il est indispensable de prendre en compte ces représentations multiples du Web, effectives ou subjectives, au même titre que l'échelle très étendue (et ouverte) des espaces considérés.

[50] Métaphore que l'on retrouve filée jusqu'à la définition du terme d'écume pour désigner l'espace des services du Web et plus particulièrement le Web social (voir à ce sujet (Rieder 2010)).

Pour les chercheurs en Sciences Humaines et Sociales (SHS) et en leur sein les SIC, le Web suscite un intérêt croissant mais soulève aussi de nombreux questionnements. La qualité des données individualisées extraites du Web est, par exemple, régulièrement mise en débat et critiquée au regard des exigences de validité, de pertinence et de représentativité des résultats obtenus et des interprétations qui en sont faites. À côté de cette critique émerge une perspective qui, d'une part, intègre ces caractéristiques comme inhérentes aux données du Web et, d'autre part, questionne les méthodes d'investigation classiques en SHS.

Il s'agit aussi d'un problème plus vaste qui porte sur le rôle et la place des SHS au sein de la société contemporaine. Le capitalisme de la connaissance (*Knowledge Capitalism*) introduit par N.Thrift (Thrift, 2005) soulève la double question de la possibilité d'une recherche empirique en SHS et de sa légitimité. Face à des acteurs de l'économie numérique qui disposent d'accès permanents et exhaustifs à ces données, les expérimentations quantitatives mais aussi qualitatives, conduites lors de recherche en SHS, courent le risque de paraitre sous-dimensionnées et d'un autre temps. L'accessibilité des données du Web dans le contexte de marchandisation est un enjeu de régulations qui fait actuellement débat[51]. Ce débat sociétal passe également par une conscientisation de la valeur associée à leur production dont les travaux sur le *digital labor* rendent compte (Cardon, Casilli, 2015). Pour les chercheurs en SHS, il est cependant urgent de ne pas attendre pour éviter un décrochage (Savage, Burrows, 2007), (Boullier, 2015). Il s'agit de s'emparer des questions épistémologiques et méthodologiques se rapportant à la recherche empirique en SHS, dans les conditions liées à la globalisation numérique. Les questions soulevées tiennent au changement d'échelle dans le rapport aux données. Dans cette situation de « déluge de données » (Boyd, Crawford, 2012), la recherche de corrélations qui permet de rendre compte d'un état de fait est privilégiée. La recherche de causalités, historiquement à l'origine de la construction de la science, est devenue maintenant plus difficile à mettre en œuvre dans ces conditions (Mayer-Schönberger, Cukier, 2013).

3.2. Une lecture socio-technique des comportements info-communicationnels

Le cadre général de mes travaux articule *deux ordres de réalité* : social et technique, selon la distinction qu'introduit Madeleine Akrich (Akrich, 1989) et sans pour autant supposer que ceux-ci soient autonomes[52] ou assujettis unilatéralement l'un à l'autre (déterminisme technique versus déterminisme social) (Proulx, 2001). Les problématiques de l'innovation, généralement attachées au couple socio-technique, qui interrogent la relation entre ces deux ordres ne constituent pas le cœur de mes travaux. En effet, l'innovation socio-technique se rapporte le plus souvent au temps court suivant l'émergence d'une pratique ou l'apparition d'un nouveau dispositif, alors que les problématiques que j'ai abordées jusqu'à présent se situent plutôt dans le temps long des comportements installés et des logiques d'usage (Perriault, 1989), (Jouët, 2000), (Proulx, 2015).

Situer socialement l'activité observée et dépasser la notion de poste de travail, m'est apparu rapidement incontournable. Cela rejoint la question fondamentale portant sur le rapport

[51] Projet de la loi pour une république numérique : http://www.economie.gouv.fr/projet-loi-numerique.

[52] Ayant leur logique propre.

instrumental exprimé dans l'interaction homme-machine. Cet élargissement problématique rend secondaire l'activité instrumentée, assujettie à une activité principale, qui s'élabore et s'établit dans un cadre plus large suivant un déterminant socio-culturel.

Le parti pris théorique supportant cet élargissement est celui d'une théorie de l'action située (Theureau, 2004), (Suchman, 2007). Ce choix théorique impose, dans ses prolongements, l'hypothèse d'une activité info-communicationnelle dans laquelle s'inscrit la rationalité opératoire de l'interaction avec les dispositifs de cette nature. Cette hypothèse rejoint les postulats théoriques actuels en SIC (Chaudiron, Ihadjadene, 2010), (Paganelli, 2012). Elle est suffisante si la finalité est de nature strictement informationnelle et si la situation ne comporte qu'un seul individu confronté isolément au dispositif. En revanche, cette hypothèse devient caduque si l'interaction trouve son objet dans des considérations autres qu'informationnelles et tout particulièrement si la médiation numérique implique d'autres acteurs.

Si l'évolution des technologies d'information communication a été rapide au cours de ces vingt dernières années, leur diffusion et leur appropriation sociale n'en sont que plus saisissantes. L'émergence d'environnements *pervasifs* ainsi que l'offre croissante de bouquets de services info-communicationnels du Web, estompent les contours des machines au bénéfice de fonctions accomplies, disponibles à tout instant et en tous lieux. L'individu isolé face à son ordinateur n'est plus une caractérisation pertinente de la situation d'interaction. Désormais, ni l'individu ni son activité ne peuvent s'envisager sans l'attribution d'un rôle ou sans l'existence de pairs contributeurs, observateurs ou évaluateurs de l'activité en cours. Selon moi, dans ce renouvellement du contexte d'interaction, la compréhension des comportements info-communicationnels ne peut plus se concevoir isolément :

- en référence à des modèles de pratiques établies *in abstracto* ;
- dans le cadre réduit d'une approche scientifique mono disciplinaire.

Ces deux points suggèrent que l'étude de l'activité info-communicationnelle médiatisée doit être mise en perspective dans un cadre d'activité socio-culturelle qui la surplombe. L'élargissement du focus à une activité socialement située étend également le dialogue interdisciplinaire que ce cadre plus général convoque. De plus, à partir des projets que j'ai conduits, il m'apparait que l'analyse des phénomènes et des comportements info-communicationnels constitue une plateforme permettant d'établir un dialogue interdisciplinaire en SHS (analyse politique et sociologie des médias par exemple).

3.3. Une approche empirique et expérimentale des comportements info-communicationnels

Fixer son attention sur l'utilisateur final a été décisif dans la démarche qui m'a conduit à étudier les interactions Homme-Ordinateur puis les phénomènes et les comportements info-communicationnels. Cette posture épistémologique impose un ancrage méthodologique fort, indispensable selon moi, pour que la réflexion théorique soit mise à l'épreuve des faits.

Il doit être considéré nécessaire de produire des observables et constituer des observations dans l'évolution des pratiques scientifiques en Sciences humaines et sociales et en Sciences de

l'information et de la communication en particulier (Courbet, 2011). Dans le cadre d'analyse retenu, les observables articulent deux programmes de recherche distincts portant respectivement sur les aspects internes et externes de l'interaction individuelle médiatisée. Le premier repose sur le paradigme instrumental et l'analyse de l'agir individuel situé. Suivant ce programme, la présence d'autres intervenants dans l'écosystème se résume à leurs contributions informationnelles. Le second repose sur le paradigme de médiation et se rapporte à la communication inter personnelle *médiée*. Pour ces deux programmes, le point de départ devient la constitution de corpus de données d'interaction et l'exploration de ces traces numériques d'activités.

Depuis quelques années, la notion de trace numérique fait l'objet d'un consensus plus ou moins tacite entre les sciences de l'ingénieur (Laflaquiere & Al., 2008) et les SHS (Perriault, 2009), (Galinon-Mélénec, Zlitni, 2013) pour désigner les données numériques associées aux actions individuelles et auxquelles peuvent accéder les systèmes informatiques. La provenance de ces données est multiple. Elles sont issues d'interactions ou produites à partir de traitements auxquels un même identifiant a été associé. Elles ne correspondent pas toujours à une intentionnalité et ne sont pas nécessairement visibles ou intelligibles. Elles relient de manière plus ou plus moins forte, légitimement et durablement, l'individu qu'elles caractérisent à un instant donné, à une personne.

Constituer les comportements info-communicationnels instrumentés en objet d'étude impose de disposer d'outils et de méthodes de recherche adaptés à leurs évolutions. Le régime quasi-permanent qui s'est instauré dans la relation personnelle aux dispositifs info-communicationnels, limite fortement l'intérêt d'observations en laboratoire. La disponibilité des profils d'abonnements, comme celle des publications ou contributions individualisées publicisés par les plateformes de services, mettent en évidence de nouvelles perspectives de traçage qu'on ne peut pas négliger (Boullier, 2015). Ces traces numériques font l'objet de nombreuses spéculations : sur ce qu'elles dévoilent ou déforment de notre identité ; sur ce qu'elles nous permettent d'élaborer et faire coexister comme identités substituées.

En tant que données, les traces entrent dans une perspective computationnelle qui répond aux attentes d'une société régie par la mesure et le discours expert (Mattelart, 2001), ainsi qu'à celles d'un marché qui mise beaucoup sur l'individualisation de masse. S'intéresser aux données de traces c'est, d'une part, aligner les perspectives d'une recherche sur les logiques d'une société et d'un marché et ainsi se donner les moyens d'évaluer (anticiper) la portée des questionnements soulevés. C'est, d'autre part, faire l'hypothèse que l'accumulation de ces marqueurs d'activités peut en restituer une vue cohérente et phénoménologique, à partir de laquelle l'interprétation dans des termes de comportements et de pratiques devient possible. Cette hypothèse est accréditée par la cohérence que les fournisseurs de services doivent maintenir, entre les dispositifs de traçage (*tracking*) qu'ils déploient et les logiques d'action des usagers, pour rendre interprétables les traces.

Le terme d'*accountability* introduit par Harold Garfinkel (Garfinkel, 1967) est aussi employé par l'informatique des réseaux, pour signifier la traçabilité en responsabilité de l'utilisateur. Pour ma

part, j'adopterai ce terme[53] pour articuler les enjeux théoriques de l'agir situé (cf. *infra*) se rapportant à l'individu avec ceux de l'économie numérique qui se situent au-delà de la définition informatique.

Les traces numériques associées aux plateformes de publications sont de plus en plus organisées en flux de messages horodatés (*timeline*), croisant les logiques de la communication interpersonnelle et de l'information indexée (*hashtags*). Mises en œuvre dans le contexte des réseaux sociaux numériques (RSN) (Facebook, Twitter, etc.), ces modalités de publication incitent à franchir un pas supplémentaire portant sur l'étude des interactions sociales et des comportements info-communicationnels collectifs.

Ces plateformes de services jouent un rôle important dans les différentes formes de mobilisations collectives qu'il s'agisse de médiatiser une cause ou une action, ou de prendre part à un événement médiatique majeur. Adressant une audience, elles sont intégrées dans des stratégies d'acteurs qui les mobilisent dans une communication multicanal.

Analyser les traces publicisées sur les canaux que les RSN organisent permet, à partir de l'étude des comportements info-communicationnels individuels (et sous réserve de leur intelligibilité), d'envisager celle de l'action collective qui sous-tend la participation militante ou partisane. Ce sont alors les phénomènes collectifs produits par les comportements individuels, qui deviennent le matériel de telles investigations.

L'approche des interactions info-communicationnelles par les traces numériques massives, suppose des choix méthodologiques, notamment celui de passer par des modèles descriptifs *implémentables* et pouvant répondre à une perspective de calculabilité. Cette perspective computationnelle rejoint les préoccupations inférentielles et prédictives des industries du Web. Par ailleurs, la formulation de modèles calculables a une vertu heuristique inductive dans la mesure où ces modèles « *assurent un rôle de médiateur entre la théorie et les données* » (Armatte, 2005).

L'empirisme que sous-tend la formulation de ces modèles intermédiaires trouve sa légitimité dans le provisoire d'un programme de recherche élargi qui vise à combler leurs défauts. Dans ce sens, la démarche empirique poursuivie contribue, selon les termes de Robert Merton à l'émergence d'une *théorie à moyenne portée* (Merton, 1997).

4. Structure du mémoire d'habilitation à diriger les recherches

L'objectif de ce mémoire est d'apporter une contribution de nature méthodologique et critique portant sur l'analyse des pratiques info-communicationnelles instrumentées suivant les logiques et dans le temps long de l'usage social. La nature de mes travaux circonscrit cette réflexion aux dispositifs et technologies issus du Web et de ses évolutions collaboratives (en abrégé Web2.0[54]). Pour mener à bien cet objectif, le mémoire se décompose en deux parties.

---

[53] Je renvoie à la lecture critique qu'en propose L. Quéré sur la pertinence sociologique du concept (Quéré, 1987). Affiner cette définition portera une partie de mon travail dans ce mémoire, notamment parce que concept proposé par Garfinkel renvoie aussi à la pratique scientifique.

[54] Je ne reviendrais pas sur le bien-fondé de cette notion à la fois catégorie et jalon de l'histoire du Web. Je renvoie à la lecture de l'ouvrage Philippe Bouquillon et de Jacob T. Matthews pour leur regard critique (Bouquillon, Matthews, 2010)

4.1 Première partie

L'objectif de la première partie est de définir un positionnement relatif à des concepts, des paradigmes, des sujets ou des objets de recherches dont l'appartenance au champ disciplinaire fait consensus. J'adopte comme postulat l'affirmation historique de l'appartenance au champ des SIC des Technologies de l'Information-Communication analogiques ou numériques.

Pour affiner cette position, je développerai la notion de *dispositif info-communicationnel* (chapitre 1) pour décrire la complexité des situations qu'engendrent la convergence et la permanence numérique (spatiale et temporelle). Largement diffusée au sens d'agencement, cette notion complexe acquiert progressivement une signification restituant la tension exprimée dans le couple socio-technique. De manière générale, je montre que l'utilisation du terme *dispositif* se diffuse dans notre discipline en se déclinant suivant les spécificités de ses champs d'études. Cependant, l'interprétation qu'on en donne reste en retrait d'une définition qu'il faut pousser plus loin afin de saisir les évolutions sociales qu'induisent la diffusion des technologies d'objets connectés et l'extension des environnements pervasifs. La notion de *dispositif info-communicationnel*, qu'il convient néanmoins d'affiner, me parait intéressante par son caractère heuristique. Ce constat plaide en faveur d'une conceptualisation de ce qui m'apparait n'être encore qu'une prénotion. Cette réflexion s'impose tout particulièrement dans le domaine du numérique ou la notion de TIC semble dépassé.

Parce qu'elle est une question naturellement associée à mes travaux, j'aborde les *approches méthodologiques de l'usage* (chapitre 2) telle qu'elles se sont développées et organisées depuis les années 1990. Il s'agit, dans une perspective renouvelée de dispositif info-communicationnel, et suivant une logique empirique associée aux traces d'usage, de relire de manière critique les fondements des méthodologies de l'analyse de l'usage dans le champ des SHS en général et des SIC en particulier. Si l'usage est largement abordé dans les sciences de l'information et de la communication, c'est essentiellement dans une perspective se référant à la *sociologie des usages*. Ce domaine frontière, dont l'intérêt est manifeste, soulève cependant la question d'une réappropriation disciplinaire des enjeux de l'usage. Je tenterai de caractériser les approches méthodologiques qui paraissent les plus représentatives et les plus adaptées pour analyser les usages et leurs évolutions dans le contexte renouvelé de dispositifs info-communicationnels. Le programme de recherche que je poursuis, m'incite tout en restant cohérent avec ses logiques, à dépasser le cadre d'étude de la sociologie de l'usage pour rejoindre celui de l'activité située. L'objectif, à l'issue de ce chapitre qui clôt cette première partie, est de disposer de l'ensemble des éléments nous permettant d'ouvrir une seconde partie portant sur les questions méthodologiques et théoriques que l'observation des pratiques effectives sur le Web soulève.

Pour aller dans ce sens et préciser la démarche empirique, j'aborde à la suite du précédent chapitre, les problématiques des *traces d'usage* (chapitre 3) comme source de connaissances sur la personne agissante. Ce chapitre envisage dans une première partie les conditions socio-techniques de l'émergence du principe de traçabilité associée à la notion d'*accountability* sur Internet. Les traces dont il est question ici sont numériques. Elles sont produites et administrées à partir des services qui structurent l'offre du Web. Ces traces d'usage, qui suivent les évolutions des utilisateurs, sont ensuite considérées en tant que données d'observation dans le cadre de recherches sur l'usage info-

communicationnel du Web mais pas seulement. Sources d'enjeux sociétaux considérables, alimentant des craintes autant que des espoirs, fondant l'économie des données, les traces numériques ont contribué au développement de techniques analytiques (*Web Usage Mining*) qu'il convient de désenchanter. Ces techniques pleines de promesses, pas toujours bien maîtrisées, bouleversent néanmoins les méthodes et les stratégies de communication commerciale et politique. Issues de l'informatique et des statistiques, les méthodes proposées, à fort potentiel heuristique, doivent faire l'objet d'une appropriation par le champ des SIC, de manière à nourrir la démarche empirique et à mettre à l'épreuve les modèles théoriques à la lumière des comportements info-communicationnels actuels. Ces enjeux ont également un prolongement social majeur. Il s'agit, alors que les *Big Data* suscitent beaucoup d'intérêt un sein de la société, de fonder une position critique face à des stratégies d'acteurs qui investissent les données et les calculs d'un pouvoir de prédiction et de mesure (Cardon, 2015).

Les données de traces ne sont certes pas les seules *input* permettant d'aborder l'usage des dispositifs, et la fouille des données d'usage n'est pas non plus la seule base méthodologique disponible. Cependant, il existe des situations d'usage ou de pratiques difficiles à appréhender sans avoir recours à l'analyse de ces traces. Ces difficultés se rencontrent dès l'échelon individuel et se généralisent à l'échelon de collectifs d'individus interagissant simultanément et produisant ainsi des phénomènes que seules les traces d'usage permettent de restituer et d'analyser. C'est par exemple le cas des phénomènes de résonnance ou *buzz* qui surviennent dans le bruissement médiatique et se propagent dans les Réseaux Sociaux Numériques (RSN). La nature contingente de ce type de phénomènes, par ailleurs massifs, renvoie à la dynamique interne des dispositifs et à la nécessité de les saisir sur le vif. Nous verrons que seule la mise en œuvre des traces d'usage et de leur analyse, permet cet ajustement aux variations de configurations et aux dynamiques internes. Davantage que la question des volumes, dont certains aspects peuvent se traiter par échantillonnage à l'instar des méthodes classiques en SHS, nous constaterons que c'est la double question de la complexité structurelle et du rapport social au temps (Rosa, 2011) qui justifie pleinement le recours à ces méthodes.

4.2. Seconde partie

Cette seconde partie s'appuie de manière explicite sur les travaux et sur les publications auxquels j'ai contribué. Je proposerai de porter une réflexion méthodologique sur l'ensemble des opérations intellectuelles et mécaniques, informatisées ou non, que la démarche que je poursuis rend nécessaire afin de soutenir un projet scientifique empirique.

Dans un premier chapitre consacré aux *approches expérimentales et dispositifs de traces* (chapitre 4), je reviendrai sur les trois contextes historiques, scientifiques et techniques dans lesquels j'ai mis en œuvre des observations expérimentales pour des dispositifs info-communicationnels. Ces trois époques espacées de dix ans chacune correspondent à l'évolution de situations d'interaction qui vont du poste de travail fixe dont l'environnement logiciel est spécialisé, aux situations mobiles et multi supports articulées autour des services du Web. La présentation des trois dispositifs d'observation met en évidence ces évolutions et leurs incidences méthodologiques.

Nous verrons que, dans mon travail, la réalisation de plateformes informatisées pour conduire des observations supervisées sur des dispositifs particuliers remplit une double fonction : celle d'une part, de rendre possible l'observation en produisant des représentations ; celle d'autre part, de formaliser tout ou partie des étapes de ce processus. La répétition des observations expérimentales dans des contextes variés a été l'occasion d'élaborer une architecture logicielle modulaire supportant une grande variété de processus d'observation. Ainsi, par raffinement successif la réalisation informatique devient l'expression d'une formalisation computationnelle de l'observation dans le contexte numérique. Cette formalisation caractérise le fonctionnement d'un dispositif dans lequel le scientifique n'est pas le seul acteur.

L'enregistrement numérique permet de s'affranchir de la synchronie de l'observation. Cette fonctionnalité, indispensable pour le travail d'analyse, exige la prise en compte de l'ensemble des informations relatives aux contributions effectives et à leurs modalités pour les différents acteurs de l'observation. Je proposerai le terme de *collection* pour désigner les éléments collectés tout au long d'un cycle d'observation. Cet objet représentationnel numérique est investi d'enjeux qualitatifs. Ces enjeux sont liés aux projets scientifiques et analytiques qui sous-tendent l'observation. Par ailleurs l'autonomisation de la collection par rapport à son contexte de création fait apparaitre des enjeux documentaires de conservation, de médiation et de diffusion. Ces différents enjeux sont intégrés tout au long de la fabrication de la collection et se reportent dans sa représentation.

Ce processus correspond à la cascade de transformations qui implique en *input* les données publicisées relatives aux traces d'usage, les données extraites des publications sur le Web, les données de *monitorings* et différents enrichissements documentaires, et qui produit en *output* les données de collections. Ainsi, envisagé le processus de collection devient un objet scientifique auquel s'attache aussi la réflexion méthodologique.

Celle-ci est abordée à la suite, dans un chapitre consacré aux *aspects méthodologiques de la collection de traces d'usage* (chapitre 5). L'objet principal de ce chapitre porte sur la réalisation de *collections*[55] numériques de données dans un cadre scientifique de recherches en Sciences humaines et sociales. La réalisation de collections intègre les ambitions et les contraintes contemporaines de projets recherche devant justifier d'une légitimité sociale autant que scientifique. Ces différentes justifications ont une incidence dans la manière d'aborder la mise en œuvre de la collecte de données ainsi que dans les choix représentationnels de leur enregistrement. Ainsi doit-on composer avec de nouvelles ambitions portées par exemple par un changement d'échelle lié aux *Big Data* et des attentes fortes exprimées dans des plans de gestion de données. Symétriquement, les conditions de mise à disposition de données par les dispositifs info-communicationnels, dans un contexte global d'industrialisation et de marchandisation de celles-ci, posent la question des limites de leur accessibilité autant que de leur validité ou de leur pertinence. Ainsi, la collection devient un objet représentationnel intermédiaire qui supporte ces différentes contraintes et contradictions tout en

[55] Pour mémoire, une collection désigne dans nos travaux un ensemble homogène, cohérent et pertinent d'unités d'enregistrements. Ce terme a été choisi en raison de son usage documentaire qui correspond à notre orientation.

devenant incontournable dans l'approche exploratoire des phénomènes info-communicationnels et médiatiques.

D'un point de vue plus général, l'interprétation des données de la collection ne peut pas s'envisager sans formuler des hypothèses sur le processus de production des traces d'usage originelles. Ces considérations demandent de sortir du cadre méthodologique pour être confrontées à celui de la production de connaissance et plus généralement aux cadres épistémologiques de la discipline.

C'est pourquoi, je propose de revenir sur les *enjeux scientifiques et disciplinaires* (chapitre 6) que soulève la réalisation de collections de données issues de traces numériques d'usage. Ce retour sans être exhaustif, met l'accent sur trois points clefs soulevés dans ce travail. En premier lieu, la référence aux technologies de l'information et de la communication prises dans leur globalité atteint ses limites. Il faut désormais aborder les pratiques info-communicationnelles au travers des dispositifs qui les constituent. En second lieu, l'industrialisation et la marchandisation des données de traçage font exister, dans les dispositifs info-communicationnels, un processus de *datafication* que l'on peut rapprocher du processus d'*informationnalisation* envisagé dans les travaux théoriques de Bernard Miège et Gaëtan Tremblay (Miège, Tremblay, 1999). Ce rapprochement constitue une proposition de formalisation dont le but est d'inscrire la notion de dispositif info-communicationnel dans un cadre théorique et une approche disciplinaire. À la suite, nous reviendrons sur les problèmes méthodologiques soulevés par les *Big Data* et leur incidence sur la production de connaissances dans les disciplines des Sciences humaines et sociales confrontées au « *déluge des données* ». C'est dans cette perspective que se développent différentes émergences, transdisciplinaires permettant une posture critique et assurant une fonction de redistribution des méthodes au sein des *Digital Humanities*. Je terminerai enfin en abordant les enjeux et les perspectives personnels tels qu'ils nous apparaissent à l'issue de ce mémoire.

Partie I

Préambule à la première partie

Le début des années 1980 constitue un nœud dans l'histoire des technologies de l'information et de la communication. Des appareils électroniques d'enregistrements procurent aux foyers, une plus grande autonomie et un contrôle individualisé de la consommation de contenus audio-visuels. Dans le même temps, la miniaturisation et l'intégration des composants rendent possible la constitution d'une offre informatique domestique. La télématique enfin annonce la possibilité d'une mise en réseaux et la généralisation de l'accès à des ressources informatiques. Dans ce foisonnement d'innovations convergentes, que l'on désigne alors par "nouvelles technologies de l'information-communication" (NTIC), les lignes de force qui apparaissent rendent compte d'intérêts économiques et politiques. Les premières synthèses publiées au début des années 1990, restituent clairement l'intérêt stratégique dans la constitution d'une offre, de la compréhension des *usages* et de l'indissociable figure de l'*usager*. Mais ces synthèses, sans s'opposer, s'envisagent différemment selon que la finalité réside dans la compréhension en aval de l'évolution des *pratiques* de communication (Jouët, 1993) ou dans l'intégration en amont, d'éléments légitimant les usages de ces NTIC dans le processus technique de conception (Mallein, Toussaint, 1994). Dans le premier cas, ce sont les *usages sociaux* qui vont être étudiés alors que dans le second, ce sont plutôt les *significations d'usage* (*Ibid.*).

Les travaux portant sur les usages sociaux des TIC relèvent à cette époque principalement de la sociologie. Un positionnement plus spécifique aux Sciences de l'information et de la communication est apparu dans les questionnements sur *la communication médiatisée* et sur les *médiations numériques* d'autant plus nettement que l'informatique se généralisait comme technologie. Suivant leurs spécificités, chacun de ces questionnements est plus large que la question des usages sociaux des TIC, et met respectivement l'accent sur les pratiques de communication ou les pratiques culturelles. Pour autant, les approches de la sociologie et des SIC ne se contredisent pas. À bien des égards, la démarcation disciplinaire sur ces sujets reste flottante et les apports réciproques. À l'intersection de différents intérêts et démarches d'analyse, l'étude des *usages* traduit un double renversement de perspective.

En premier lieu, reconnaitre à l'usage des outils de médiatisation et de médiations numériques une dimension sociale, c'est au travers de leur étude, se donner l'ambition d'en interroger le sens et la portée collective ; d'articuler, comme le souligne Serge Proulx, « *les significations sociales et le développement des usages aux mutations sociales en cours* » (Proulx, 2015, p3).

En second lieu, les pratiques des machines à communiquer ne sont plus considérées comme entièrement déterminées par leur finalité ni par les prescriptions qui accompagnent leur utilisation. L'usage introduit une part d'indétermination dans l'objet technique soulignant l'importance de l'observation de terrain au risque, comme le souligne Yves Jeanneret, « *de prendre pour réalité des pratiques ce qu'on en voit* » (Jeanneret, 2009, p4).

Ces deux hypothèses ont été structurantes dans l'orientation de nos travaux jusqu'à ce que le développement des technologies numériques et des réseaux nous[56] fasse percevoir les limites d'une approche dont les catégories sont devenues inadaptées. D'une part, les objets et les artefacts communicants permettent d'accomplir un très grand nombre de fonctionnalités qui ne sont plus distinctives d'objets ni d'usages. D'autre part, les lieux et les temps ne constituent plus des cadres d'activités différentiées mais s'enchevêtrent dans un continuum de pratiques.

Ce constat a une conséquence épistémologique qui nous conduit à reprendre à notre compte le concept de *dispositif* afin de l'adapter dans le contexte de pratiques info-communicationnelles. L'objectif n'est pas seulement l'utilisation d'un terme mieux adapté pour décrire la réalité d'observations des usages sociaux, mais de changer de paradigme en nous appuyant sur le postulat de la nécessité structurelle de dispositifs de communication au sein de la société. Il s'agit alors moins d'étudier l'usage isolé d'un système informatique complexe que de comprendre comment ce système associé à d'autres fait dispositif. La complexité structurelle que traduit le choix du terme dispositif fait écho à de nombreux travaux en SIC. L'évolution que nous soulignons, ne peut pas être considérée comme une intuition isolée. Nous sommes les témoins privilégiés d'une transformation numérique de la société. De nombreux autres travaux, notamment en SIC, ont recours à cette notion et en soulignent la pertinence contemporaine accrue. Il est donc opportun de faire l'état de ces réflexions tout en soulignant l'importance pour nous d'une opérationnalité du concept, c'est-à-dire de la possibilité de lui associer une méthode d'investigation empirique.

La démarche empirique en SIC, qui a prévalu dans l'investigation des usages, s'est appuyée sur des méthodes classiques d'enquête et d'observation pratiquées dans les Sciences sociales. Dans nos travaux nous privilégions les complémentarités entre les méthodes dès lors qu'elles sont pertinentes et applicables. Les limites que soulève l'étude des usages sociaux nous ont conduit à mettre l'accent sur l'adaptation nécessaire des outils et des méthodes de l'observation.

En effet, l'observation expérimentale des pratiques effectivement libres ou scénarisées s'est essentiellement développée dans des cadres et des conditions assimilables au laboratoire. Les raisons tenaient autant à la praticité du déploiement de l'instrumentation d'observation que des aspects statiques et monolithiques du poste de travail (l'ordinateur) tel qu'il pouvait s'envisager à l'époque. Dans un tel contexte, ce sont les interactions se déroulant à la surface du système, précisément sur l'écran et les périphériques de l'ordinateur qui nourrissent empiriquement les études. Les résultats de telles observations ont pu servir une compréhension sémio-pragmatique ou ergonomique des interfaces graphiques mais guère au-delà.

L'ambition dialogique, multimodale et multimédia alliée à une complexité fonctionnelle croissante, supportée par une architecture logicielle fortement modularisée et distribuée, appelle une compréhension plus fine dépassant la confrontation homme-ordinateur. Pour cela, l'approche technique des *médiatisations* et des *médiations numériques* résultant de l'évolution de l'offre technologique, nécessite de pénétrer plus profondément dans le système numérique, d'en relever

[56] Dans la suite de ce mémoire, je quitte l'expression personnelle du "je" pour le "nous" d'auteur.

les articulations fonctionnelles et les logiques internes. Le système informatique ne peut plus être seulement abordé selon sa composante matérielle et suivant les manifestations audiovisuelles de son fonctionnement. Lorsque cela est possible, les états et les représentations internes, associés à la conduite d'interactions avec les utilisateurs ainsi que la circulation de l'information entre les composants logiciels doivent également être pris en compte afin de compléter les descriptions et les représentations analytiques.

Cette proposition repose sur une double hypothèse représentationnelle. Tout d'abord, que les états des différents modules puissent être caractérisés au travers de leurs productions informationnelles. Ensuite, que la communication entre les composants soit susceptible de restituer avec une précision suffisante les logiques internes et fonctionnelles des systèmes informatiques en fonctionnement. Nous proposons comme corolaire à cette hypothèse que les dispositifs s'appuyant sur ces systèmes informatiques peuvent de cette manière être *tracés*. Ce raisonnement trouve sa légitimité dans la cybernétique. L'hypothèse implicite consiste dans le fait que la communication entre les parties corresponde aux activités de coordination ou de régulation de l'activité programmée du système informatique ou du dispositif. Nous désignons par *paradigme computationnel* cette réalité opérationnelle qui confère aux traces numériques produites, une pertinence représentationnelle. Nous assumons le postulat que ce paradigme n'est pas perverti, autrement dit que la trace est l'empreinte de l'activité. Partant de là, les traces produites dans les conditions d'usage des systèmes numériques vont retenir notre attention comme sources d'information pour l'étude des dispositifs info-communicationnels.

CHAPITRE 1
Dispositif info-communicationnel

Le terme *dispositif* désigne, dans l'espace technique, la manière dont sont agencées les différentes parties d'un mécanisme pour répondre à un but précis. L'usage militaire de ce terme fait ressortir l'intention stratégique qui sous-tend la mise en œuvre d'un dispositif mais aussi l'ensemble codifié des interactions humaines qui en conditionne la mise en œuvre. Ces deux conceptions traduisent l'utilisation ordinaire du terme. Une définition synthétique et contemporaine est donnée par Philippe Zittoun qui décrit le dispositif comme étant « *d'abord un assemblage intentionnel d'éléments hétérogènes répartis spécifiquement en fonction d'une finalité attendue*. » (Zittoun, 2013) Il se dégage de ces premiers éléments de définition une visée fonctionnelle, mécaniste, associée à un degré certain de complexité organisationnelle et technique.

Dans nos publications, ce terme apparaît à de nombreuses reprises. Certes, on peut y voir une facilité d'emploi d'un terme qui entretient une grande proximité avec les discours réflexifs produits au sein d'une société dont on dit qu'elle se technicise[57]. Évoquer des dispositifs techniques dans le champ des Sciences de l'Information de le Communication (SIC) n'a d'ailleurs rien de très original en soi. Il suffira pour s'en convaincre de consulter les résultats de l'analyse de la base de données bibliographiques des thèses produites dans la discipline (cf. § 3.1).

Cependant, le terme "dispositif" recouvre deux types de considérations dans nos travaux. Le premier se rapporte aux ressources socio-techniques mobilisées dans les pratiques sociales des technologies de l'information communication, à leur agencement et aux configurations[58] qui les conditionnent. Le second concerne l'instrumentation expérimentale nécessaire à la compréhension de ces pratiques info-communicationnelles. Ces deux cas renvoient à des logiques différentes. Dans le premier cas, l'approche est celle d'une logique analytique et critique (Appel, Heller, 2011). Dans le deuxième cas, il s'agit d'une logique descriptive d'agencement telle qu'elle est soulevée par Violaine Appel et Thomas Heller (*Ibid.*). Ces deux logiques conduisent à la production de méta discours qui se différencient assez clairement.

- Dans le premier espace, l'objectif est d'une part, de circonscrire le périmètre concerné par le dispositif et d'autre part, de saisir la complexité des mises en tension et d'interactions

[57] Argument analogue à celui développé par V. Appel, H. Boulanger et L. Massou dans le chapitre introductif (Appel, Boulanger, Massou, 2010) en référence à la publication de G. Leblanc dans la revue Hermès (vol. 25/3) consacrée également à la notion de dispositif (Leblanc, 1999). Affirmation également reprise par Y. Jeanneret (Jeanneret, 2005).

[58] Nous reviendrons sur l'emploi de ces deux notions en conclusion de ce chapitre.

entre les espaces sociaux et techniques. Dans cette perspective, le dispositif apparaît comme étant par nature évolutif, simultanément agi et agissant ;

- Dans le second espace, le dispositif est davantage envisagé comme un agencement circonstanciel, organisé et maîtrisé par la finalité expérimentale. Les présupposés de l'expérimentation scientifique (stabilité et neutralité) ne peuvent que se heurter aux caractéristiques du dispositif tel qu'il est envisagé dans l'espace précédent.

Il serait commode de considérer que la distinction de nature ou de niveau (méta ou non) des discours produits sur ou dans chacun des espaces précédents est fondée sur le fait qu'ils se rapportent à des problématiques différentes. Mais cette perspective n'est pas tenable si l'on s'en réfère à l'unité du cadre conceptuel que la démarche empirique et expérimentale introduit par l'objet d'étude. En effet, l'expérimentation est supposée de nature écologique, c'est-à-dire se fondant de manière (la plus) neutre dans l'espace technique et social des pratiques ordinaires. Dans ces conditions, le dispositif expérimental visant à assurer la lisibilité du dispositif observé doit envelopper celui-ci afin d'en restituer les flux informationnels et documentaires autant que les interactions qui se développent en son sein. Les sondes, capteurs, éléments d'enregistrement plongent dans le dispositif et dans son environnement. Il apparaît difficile, en toutes circonstances, d'éviter toutes interactions entre les deux dispositifs. Des interfaces sont nécessairement organisées (structures de données, métadonnées, etc.) et des frictions se produisent dans la mise en œuvre expérimentale (ralentissement, saturation, etc.). Dans ces conséquences, le dispositif d'observation participe effectivement au dispositif observé. Même si les ancrages dans ce dispositif ne se traduisent pas par des phénomènes signifiants pour les acteurs impliqués, l'enveloppement idéal revient à unifier les éléments constituant des deux dispositifs.

En outre, cette unité conceptuelle apparaît d'autant plus fondamentale dans la mesure où le paradigme de *performance*[59], qui sous-tend les dispositifs contemporains d'information-communication, n'est pas dissociable de celui du traçage[60] avec lequel il se confond.

L'organisation de l'interactivité des dispositifs dans les interfaces utilisateur repose sur des *schèmes d'actions*[61], lesquels doivent être lisibles pour l'utilisateur final (exigence d'usabilité) et, dans le même temps, permettre d'assurer la traçabilité et au-delà l'enrichissement des profils individuels.

C'est dans ce sens, que nous envisageons le terme d'*accountability*[62] inspiré des travaux de H. Garfinkel (Garfinkel, 1967). La mise en œuvre du principe d'*accountability* se traduit par le fait que le dispositif info-communicationnel, objet d'observation de notre part, tend aussi à remplir en lui-même une fonction d'observateur.

Il en découle les questionnements suivants :

---

[59] Dans ce contexte, le terme de performance s'exprime moins en référence à l'exécution d'une tâche de l'utilisateur, qu'en référence à la captation maximisée de son activité sur un site, d'un point de vue marketing ou commerciale : parcours ciblé, transformation en acte d'achat d'une consultation, etc.

[60] Voir chapitre 3.

[61] Voir chapitre 2, §1

[62] Voir chapitre 3, §2.1

- Dans une approche empirique, comportementale et phénoménologique des usages info-communicationnels, est-il possible de concevoir un cadre théorique cohérent supportant l'articulation des deux espaces de discours décrits précédemment ?
- Est-il possible, de faire porter à la notion de dispositif ces multiples niveaux de sens et d'assurer une fonction articulatoire dans l'espace méta discursif ?

Ce premier chapitre ouvre le débat en recherchant dans l'historique de la notion de dispositif, de son emploi dans les Sciences humaines et sociales et de son actualité dans les Sciences de l'information et de la communication les clefs conceptuelles qui lui sont associées.

Il s'agit pour nous de mobiliser ou d'actualiser un concept susceptible de rendre compte, au-delà de l'agencement matériel et technique, d'une intériorité dans laquelle les technologies de l'information-communication (TIC) ainsi que leurs usagers sont saisis et se mettent en tension. Nous suivrons Jean-Pierre Meunier lorsqu'il évoque l'éclairage théorique et pratique que le concept de dispositif produit lorsqu'on le rapproche du concept de communication et des schémas théoriques qui s'y affrontent (Meunier, 1999). La question que l'on peut envisager, alors, porte sur l'apport d'une notion telle que *dispositif info-communicationnel* au regard de celle de TIC.

Ces questionnements reportent sur la notion de dispositif les éléments d'une réponse épistémologique et paradigmatique. Dans ce chapitre, nous cherchons à apporter les éléments d'une construction théorique et opératoire du concept de *dispositif info-communicationnel,* lequel apparaît comme essentiel dans le cadre de notre travail, quand bien même le terme de dispositif serait «*omniprésent*» (Demaizière, 2008) voire «*galvaudé*» (Gavillet, 2010) dans le champ des TIC et des Sciences de l'information et de la communication.

1. Qu'est-ce qu'un *dispositif* ?

Le concept de dispositif fait l'objet d'une abondante littérature en Sciences humaines et sociales. On peut s'interroger, comme le proposent Jean-Samuel Beuscart et Ashveen Peerbaye, sur les raisons de l'usage extensif actuel de cette notion dans les SHS et sur le fait qu'elle puisse n'aboutir, en définitive, qu'à un cadre théorique consensuel minimal, appelant à être précisé dès lors qu'il est mobilisé (Beuscart et Peerbaye, 2006).

Au sein des Sciences humaines et sociales, ce concept marque une singularité française difficilement traduisible et peu exportée (*Ibid.*). La traduction anglaise de *device* ne nous est du reste jamais apparue satisfaisante pour porter le sens de dispositif info-communicationnel qui n'accepte pas d'équivalent en anglais.

1.1. L'apport de M. Foucault au concept de dispositif

Lorsqu'il est défini dans les publications scientifiques, le concept de dispositif est très systématiquement associé aux travaux de M. Foucault. Ce terme apparaît et se définit de façon progressive dans son œuvre. Il y recourt pour décrire les moyens institués régissant l'organisation spatiale et discursive d'une société contemporaine afin de répondre à ses besoins de surveillance et de contrôle sur ses membres. Entreprendre l'historique de ce concept nous conduit en suivant les

nombreux jalons bibliographiques qui le pointent, vers son livre *Surveiller et punir - naissance de la prison* publiée en 1975 (Foucault, 1975). Retrouver des éléments antérieurs à cette référence bien balisée est plus difficile. Judith Revel, suggère une piste dans son *dictionnaire Foucault,* à partir de la préface rédigée par Michel Foucault à l'édition américaine de l'œuvre de Gilles Deleuze et Félix Guattari *Anti-Oedipus : capitalism and schizophrenia*[63] (Revel, 2008, p41). Une lecture attentive de ce texte fait effectivement apparaître le terme de *dispositif* relevé par M. Foucault et mis en avant parmi les notions « *en apparence abstraites de multiplicités* » que les auteurs mobilisent (Foucault, 1994, p134). Les formulations associées à ce terme empruntent à l'œuvre introduite tout en s'inscrivant parfaitement dans la conceptualisation que M. Foucault est alors en train d'élaborer. On peut faire l'hypothèse que cette proximité intellectuelle a pu se manifester dès la première parution de l'ouvrage en français (1972), engageant M. Foucault à théoriser ce concept. Au-delà de la recherche d'une origine, ce qui frappe dans cette préface c'est qu'elle reprend à son compte les thèses du texte comme justificatif d'une approche globale dont est issue la formulation en dispositifs « *Préférez ce qui est positif et multiple, la différence à l'uniformité, les flux aux unités, les agencements mobiles aux systèmes »* (Foucault, 1994, p135).

Le chapitre consacré par Isabelle Gavillet à l'usage galvaudé du concept de dispositif est aussi une mise en garde quant à la difficulté de saisir le concept dans la pensée de M. Foucault (Gavillet, 2010). Comme elle le souligne, ce terme n'apparaît que tardivement dans l'œuvre de M. Foucault. À ce sujet, Michel De Certeau évoque les déplacements conceptuels et l'absence de stabilité dans le langage de M. Foucault[64] pour traduire ce qu'il désigne par « *procédures de la "surveillance"* » (Certeau, 1990).

Il ne faut donc pas considérer ce concept comme figé dans la pensée de M. Foucault mais comme le support d'une élaboration probablement non achevée qui intègre, chemin faisant, des concepts précédemment élaborés dont l'épistémè (en définitive un dispositif/configuration de savoirs) constitue un exemple. C'est ce dont atteste l'auteur en 1977 dans l'entretien qu'il accorde à la revue *ornicar?* Soulignant alors qu'il à la volonté de dépasser, avec la notion de dispositif, le cadre strictement discursif de l'épistémè pour y intégrer des éléments non-discursifs (Foucault, 1994, p301). Le choix du terme dispositif semble d'ailleurs répondre, chez cet auteur, à la nécessité d'adaptabilité et d'évolutivité de ce à quoi il fait référence. Cette élasticité référentielle permet à M. Foucault de faire supporter à la notion de dispositif toute l'hétérogénéité des éléments constitutifs : «*…des institutions, des aménagements architecturaux, des décisions réglementaires, des mesures administratives, des propositions philosophiques, morales, philanthropiques. Bref, du dit aussi bien que du non-dit…*» (Foucault, 1994, p299). Cet extrait, très souvent mis en avant[65], ne résume et n'épuise pas le concept de dispositif chez M. Foucault. Il permet cependant de saisir la portée de la (ré)unification qu'opère le concept

---

[63] Edition Viking Press New-York pp XI-XIV. La publication française du texte date de 1972 (Éditions de Minuit).

[64] «*Michel Foucault multiplie les synonymes, mots danseurs, approches successives d'un impossible nom propre : "dispositifs", "instrumentations", "techniques", "mécanismes","machineries",etc.*». (Certeau, 1990)

[65] Citation dont I. Gavillet dit qu'elle est source de malentendu et de biais méthodologique car coupée du contexte de sa définition (Gavillet, 2010).

en faisant coexister dans le champ scientifique « *des entités traditionnellement considérées comme inconciliables*. » (Peeters, Charlier, 1999).

La structure de *Surveiller et punir*[66] semble, en partie, guidée par une approche historique de M. Foucault ; ce qui conduit à penser (compte tenu du titre également) que l'institution prison (dernière partie) en tant que moyen de punition est l'aboutissement d'une élaboration permanente portée par deux préoccupations du pouvoir qui inversent l'ordre du titre : la punition[67] et la surveillance. Cette séquence se double d'une autre, qui est celle de la discipline qui s'élabore en partie sur la crainte (punir) et qui suppose dans sa réalisation la surveillance hiérarchique. La prison devient alors un espace disciplinaire par excellence. Le concept de dispositif n'est pas explicitement posé dans l'ouvrage ; il est même quasiment subliminal. Dans *surveiller et punir*, le concept de dispositif se construit au fil de l'exposé, non pas comme une définition mais comme un cadre qui se construit. Même si la partie surveillance et plus précisément le chapitre consacré au *panoptisme* apportent les éléments les plus explicites sur la notion de dispositif, il ne faut pas négliger les autres chapitres. La référence au corps est ainsi manifeste dans l'ensemble de l'œuvre comme lieu d'exercice du dispositif. Cette restriction au corps est aussi repérée par André Berten qui constate qu'ainsi l'esprit est affecté par l'extérieur (le corps) et non de l'intérieur. De ce rapport entre intérieur/extérieur, il conclut alors que «…*le dispositif produit de la subjectivité mais n'est pas produit par la subjectivité*. » (Berten, 1999, p35).

Un an après, la publication de *La volonté de savoir,* premier volume de *Histoire de la sexualité*, est une autre occasion de suivre l'évolution du concept décliné en tant que *dispositif de sexualité* (Foucault, 1976). S'il faut fixer une origine, ce livre nous parait être plus pertinent que le précédent. Il nous semble qu'on privilégie *Surveiller et punir* dans ce rôle parce qu'il fait apparaître dans le chapitre *panoptisme* les principes architecturaux de la surveillance naturellement assimilables au dispositif. Ce référentiel spatial et technique est moins présent dans *La volonté de savoir*. Pourtant une partie entière de l'ouvrage, découpée en 4 chapitres est consacrée à un dispositif spécifique *de sexualité*. Bien que le titre soit explicite, à nouveau, aucune définition de dispositif (spécifique ou non) n'est posée. Celle-ci ne le sera véritablement qu'un an plus tard (1977), à l'occasion de l'interview dans la revue *Ornicar?* et du questionnement explicite d'Alain Grosrichard[68] (Foucault, 1994, pp.206-329) Comme dans le précédent texte, ce n'est que dans la lecture que l'on peut en comprendre l'intention de l'auteur. Ce qui apparaît tout d'abord, c'est que les dispositifs en tant que tels, ne constituent pas l'enjeu de la publication. La question qui traverse les deux livres, forts homogènes par ailleurs, est celle des rapports de pouvoir. Dans *la volonté de savoir*, la notion de pouvoir est précisée (voir à ce sujet les pages 121-129) (*Ibid.*). On comprend alors que le *dispositif* est le lieu d'expression de ce pouvoir qui n'apparaît pas comme surplombant, mais diffus et mouvant, en perpétuelle redéfinition

---

[66] Nous n'entrons pas dans l'explication de l'ouvrage qui appelle plusieurs niveaux de lecture (disciplinaires) et qu'il faut saisir : en fonction du contexte d'une société qui évolue, au regard d'une institution carcérale en crise, et en report avec les engagements de l'auteur notamment dans le groupe d'information sur les prisons (GIP).

[67] Les deux premières parties de l'ouvrage reflète une chronologie historique et une élaboration technique que marque la transition de la partie 1: supplice (peine de mort) à la partie 2: la punition (emprisonnement).

[68] "*Le jeu de Michel Foucault*" entretien avec D. Colas, A. Grosrichard, G. Le Gaufrey, J. Livi, G.Miller, JA Miller, C. Millot, G. Wajeman) *Ornicar?* Bulletin périodique du champ freudien,n°10, juillet 1977, pp. 62-93.

dans les rapports de forces définissant le dispositif. La relation forte qu'il fait apparaître entre pouvoir et savoir constitue le socle de la définition qu'il donnera en 1977.

Au-delà de l'unité, ce que cherche à mettre en évidence M. Foucault, c'est « *la nature du lien* ». Le dispositif devient alors « *le réseau qu'on établit entre ces éléments* » (Foucault, 1994, p299). La structure de réseau qui est décrite n'affecte pas des positions précises, ni des rôles ou fonctions.

J. Revel estime que c'est l'émergence du concept de dispositif dans son discours qui pousse M. Foucault à un glissement méthodologique dans son analyse du pouvoir, l'amenant à se concentrer sur les mécanismes de la domination (Revel, 2008). C'est bien la nature stratégique des dispositifs qu'il convient de souligner. « *Le dispositif est donc toujours inscrit dans un jeu de pouvoir, mais toujours lié aussi à une ou des bornes de savoir, qui en naissent mais, tout autant les conditionnent. C'est ça le dispositif : des stratégies de rapports de forces supportant des types de savoir et supportés par eux.* » (Foucault, 1994, p300).

Il ressort une certaine circularité de cette définition d'un dispositif fonctionnel saisi sur le vif. Elle traduit un principe de régulation interne qui assure l'adaptation du dispositif dans le temps. Ainsi le dispositif, originellement établit et orienté dans une stratégie spécifique, dispose des moyens de son devenir. Pour M. Foucault, ce devenir est canalisé par un double processus : « *de surdétermination fonctionnelle d'une part, puisque chaque effet, positif ou négatif, voulu ou non voulu, vient entrer en résonance, ou en contradiction, avec les autres, et appelle à une reprise, à un réajustement, des éléments hétérogènes qui surgissent çà et là. Processus de perpétuel remplissement stratégique, d'autre part.* » (Foucault, 1994, p299).

Dans l'évolution du concept chez Michel Foucault, Isabelle Gavillet, se fondant sur le recueil *Dits et écrits II 1976-1988*[69]*,* relève trois périodes soulignant l'évolution de la problématisation associée au concept (Gavillet, 2010). Il est effectif que les réponses qu'apporte M. Foucault à l'occasion de son interview dans la revue *Ornicar?* instaurent un changement de perspective (cf. *infra*). Ce qui a été introduit initialement par le dispositif l'assujettit au pouvoir. Ce qu'établit l'interview c'est une autonomisation du dispositif en tant que technique. Un troisième temps est proposé par I. Gavillet qui voit dans la suite de ces travaux un possible retour sur la théorie du pouvoir et de ses finalités. Mais l'argument, s'il peut être discuté dans une analyse de l'auteur, n'affecte pas fondamentalement, nous semble-t-il, les mécanismes de pouvoir qu'il a mis en évidence.

M. De Certeau critiquera la définition du dispositif chez M. Foucault qu'il juge trop enfermante. Pour M. De Certeau, il existe en effet, des interstices ouverts à «*…d'autres procédures infinitésimales, qui n'ont pas été privilégiées par l'histoire et qui n'en exercent pas moins une activité innombrable entre les mailles des technologies instituées.* » ce qu'il appellera des « tactiques » et une possibilité individuelle de liberté (Certeau, 1990)[70]. Pour M. Foucault, cette liberté ne peut s'acquérir que dans l'affrontement. La position du dispositif au regard du *sujet* [71]semble chez M. Foucault demeurer une extériorité, ce qui n'interdit pas d'envisager une incorporation des contraintes que celui-ci exerce sur le sujet.

---

[69] (Foucault, 1994)

[70] Alors que M.Foucault considère la possibilité de libération dans la résistance contre le dispositif.

[71] Ici la personne qui "subit" le dispositif.

1.2. Les reprises théoriques

C'est bien parce que la notion de dispositif apparaît, *in fine*, comme un élément pivot de la pensée de Michel Foucault, que la question : « *qu'est-ce qu'un dispositif ? »* va être posée par Gilles Deleuze (Deleuze, 1989) contribuant ainsi à la diffusion du concept. Des travaux conduits en philosophie montrent néanmoins que le concept ne fait pas véritablement écho dans la pensée de G. Deleuze. Son intérêt est davantage celui d'un hommage et de regards croisés que les deux philosophes ont tramé entre leurs travaux[72]. Le prolongement deleuzien du concept de dispositif est abordé dans la toute récente remise en perspective du concept de dispositif qui vient d'être produite par Laurence Monnoyer-Smith dans le contexte de la parution du « *Manuel d'analyse du Web* » dirigé par Christine Barats (Monnoyer-Smith, 2016)[73]. Cette auteure considère que G. Deleuze « *surinterprète* », dans une voie et un questionnement qui sont les siens, des aspects du dispositif auxquels M. Foucault n'a pas semblé porter d'intérêt, comme le changement de nature des dispositifs (*Ibid.*, p22).

Dans un opuscule récent consacré à ce même questionnement, Giorgio Agamben revisite exclusivement la définition de M. Foucault. L'époque n'est plus la même, le téléphone portable est devenu le symbole de la permanence et de l'emprise technique qui modifie nos modes d'existences en exerçant dans ce cas particulier une forme d'impératif de connexion auquel il devient quasiment impossible d'échapper. C'est dans ce contexte que G. Agamben propose une relecture du concept en lui attribuant une généalogie théologique et philosophique. Cette relecture le conduit à une définition qui met l'accent sur un processus de subjectivations multiples (Agamben, 2007, p42). Il en vient alors à poser sa propre définition comme étant : «…*tout ce qui a d'une manière ou d'une autre, la capacité de capturer, d'orienter, de déterminer, d'intercepter, de modeler, de contrôler, d'assurer les gestes, les conduites, les opinions et les discours des êtres vivants.* » (*Ibid.*, p31). Cette définition le conduit à ne retentir que deux classes d'entités dans le monde : d'un côté, les êtres ou les substances, de l'autre, les dispositifs. Le sujet est alors l'entre deux, découlant de la mise en relation de l'être avec les dispositifs subjectivants dans lesquels il est pris. Le propos de G. Agamben est, en définitive, une critique de la société contemporaine capitaliste dans laquelle les dispositifs technologiques opèrent selon lui une double action de subjectivation et de désubjectivation ne conduisant plus à la recomposition d'un nouveau sujet. Selon l'auteur, les sujets deviennent alors insaisissables et l'exercice du gouvernement impossible sauf dans une parodie politique.

---

[72] Voir par exemple le texte *Désir et plaisir :* paru dans Le magazine littéraire, n°325, octobre 1994, et repris dans G. Deleuze, *Deux régimes de fous*, Minuit, 2003. Version en ligne : http://www.multitudes.net/Desir-et-plaisir/

[73] Nous reviendrons en conclusion de ce chapitre sur l'intérêt que L. Monnoyer-Smith dégage de la notion de dispositif et de ses prolongements deleuziens dans le contexte de l'étude du Web.

2. Dispositif socio-technique[74]

Spécifier la notion de dispositif en lui adjoignant un qualificatif revient, à partir d'une racine commune, à constituer une classe de dispositifs et par là même à entrer dans une logique catégorielle. L'expression *dispositif socio-technique* est apparue au tournant des années 1990. On en trouve une trace discrète dans un article écrit par Dominique Boullier portant sur les messageries électroniques, sans que les spécificités de la classe ne soient spécifiquement établies (Boullier, 1989). Pour des dates antérieures à 1989, une recherche sur Web (Google Scholar) ou sur la plateforme HAL de cette expression, ne donne aucun résultat[75]. Cette formulation hybridée semble s'être installée dans la continuité et la légitimité des travaux se rapportant à l'innovation socio-technique dans lesquels les concepts de *système socio-technique* et *de dispositifs techniques* sont mobilisés simultanément et se répondent (Akrich, 1987), (Akrich, Callon, Latour, 1988), (Akrich, 1989).

Jean-Samuel Beuscart et Ashveen Peerbaye trouvent dans la sociologie de l'innovation impulsée par Madeleine Akrich, Michel Callon, et Bruno Latour, une filiation avec la théorie foucaldienne du dispositif, notamment dans le choix d'une terminologie spécifique, distincte des concepts de *système* et de *structure* dont la connotation historique affecte l'usage[76] (Beuscart, Peerbaye, 2006). Ils évoquent cependant une entrée discrète « *en contrebande* » de ce concept dans le vocabulaire de la sociologie des sciences et techniques.

Il semble que la notion de dispositif socio-technique n'apparaît qu'en 2012 dans l'article *Humains, non humains : morale d'une coexistence* publiée par Michel Callon et Arie Rip. On trouve, tout au long de cet article, un ensemble de notions reliées au concept de dispositif socio-technique qui en restitue, au final, une définition tout à fait intéressante. Nous prendrons ce document comme référence emblématique dans la mise en perspective de la définition de dispositif issue de l'approche de M. Foucault et de celle des auteurs associés à la sociologie de la traduction aussi désignée théorie de l'acteur-réseau[77] ou *Actor Network Theory* (ANT).

L'article dont il est question part de la définition d'une norme néerlandaise de santé publique et environnementale fixant une valeur seuil de concentration dans l'air d'une substance chimique nocive. Cette valeur est établie en fonction d'une acceptabilité sociale du risque de cancer encouru. Les auteurs construisent en premier lieu, la notion de *norme socio-technique* dans laquelle se mêlent des catégories d'humains et de non-humains. Dans ce texte, la notion de norme socio-technique désigne « *un ensemble de règles et de prescriptions concernant la nature et la forme des rapports entre certaines catégories d'humains et de non-humains.* » (Callon, Rip, 1992, p141)

M. Callon et A. Rip soulignent que l'importance de ces normes hybridées, mêlant différentes catégories d'êtres (humains, non-humains) ou d'artefacts techniques, va croissant et « *qu'elles tendent*

---

[74] Socio-technique (M. Akrich) ou sociotechnique (M. Callon) ? Choix difficile. Pour une question d'harmonisation des formes graphiques et aussi historique, nous privilégions le trait d'union qui souligne l'articulation et la tension.

[75] On trouve effectivement des textes évoquant simultanément des dispositifs techniques et des contextes socio-techniques mais aucun forgeant le terme de dispositif socio/technique (quelle qu'en soit la graphie).

[76] Une remarque analogue justifie pour M. Foucault le choix de dispositif.

[77] L'intitulé de l'article incite d'ailleurs à cette mise en perspective programmatique.

à se substituer à un univers de règles ou de conventions qui seraient soit "purement" sociales, soit "purement" techniques. » (*Ibid.*, p141). Ces auteurs décrivent ensuite, le rôle de l'expert comme celui d'un médiateur qui assure le double travail de traduction et de négociation des résultats scientifiques dans une norme socio-technique qui devient alors un compromis « *socialement viable ou acceptable* ». Ce passage par l'acteur expert leur permet de définir l'expertise comme « *le dispositif permettant d'établir, alors qu'on ne le connaît pas, la carte des gradients de résistance* ». Ce dispositif permet d'identifier le point d'équilibre qui est alors « *un assemblage d'éléments hétérogènes considérés comme suffisamment robuste pour avoir une certaine stabilité : ce que l'on sait (ou croit savoir) et ce que l'on décide (ou croit décider) se trouve aligné et se renforce mutuellement.* » (*Ibid.*, p147).

Dans une démarche de conceptualisation et afin de rendre compte des situations complexes dans lesquelles une large part de la société est conviée (ou se convie) au débat, les auteurs étendent la notion d'expertise à « *l'ensemble des mécanismes et des dispositifs qui permettent d'aboutir à un alignement durable »,* dont les normes socio-techniques découlent. L'expertise est alors un processus, qui se déploie au sein de ce qu'ils nomment *« un forum hybride »*. Celui-ci est constitué, dans une situation donnée (*i.e.* une norme) par des acteurs qui interagissent et nouent des réseaux d'alliances en fonction d'enjeux et d'intérêts qu'ils partagent ou qui les opposent contextuellement.

La notion de *forum hybride* traduit la très forte interpénétration des acteurs et des débats au point de rendre impossible toute distinction de rôles. Les forums hybrides échappent à l'emprise des pouvoirs (souverain, juridique, scientifique)[78]. M. Calon et A. Rip leur associent un fonctionnement généralement confus.

Il s'agit, selon ces auteurs du «…*creuset où se transforment et s'adaptent tout à la fois la société, la technique et nos savoirs sur la nature »* (*Ibid.*, p151). En particulier, les normes socio-techniques s'y élaborent dans une négociation simultanée « *des savoirs, de l'identité de certains acteurs*[79] *et des procédures à suivre pour établir les normes* ». Dans ce cadre, ils (re)définissent l'expertise comme étant « *l'ensemble du dispositif socio-technique qui crée les conditions de la production de l'accord, c'est-à-dire l'alignement entre les trois pôles* (les technosciences, le droit et des réglementations, le monde sociopolitique et économique)*, et il n'est plus question de le limiter à un groupe particulier, voire aux seuls acteurs humains.* » (*Ibid.*, p155).

Leur approche descriptive et l'effort de généralisation proposé par ces auteurs inspirent notre démarche d'analyse du fonctionnement d'une plateforme de services telle que Twitter. Les notions de forum hybride, d'alignement négocié et normatif rejoignent nos observations de Twitter relatives aux événements médiatiques politiques ou sportifs.

En comparaison, il nous semble que la notion de dispositif employée dans le contexte socio-technique ne se situe pas sur le même plan que le concept développé par M. Foucault. Ce dernier parait plus proche du concept de *norme socio-technique*, celle-ci agissant sur la société dans son entièreté. La norme édicte par ailleurs, des comportements et servitudes que l'on peut rapprocher de logiques de contraintes et de pouvoir, alors que M. Calon et A. Rip détachent de la notion de

---

[78] Cette définition n'est pas sans lien avec la problématique de l'espace public et de son instanciation dans les espaces numériques, analyse à laquelle nous convient nos travaux sur les événements médiatiques (cf. partie 2).

[79] La question de l'identité n'est pas ici très claire. C'est peut-être la figure de l'expert qui s'y trouve évoquée.

pouvoir celles de forum hybride et de forces qui le traversent. Sans nier les enjeux de pouvoir qui s'inscrivent dans les dispositifs, ceux-ci n'en sont pas les seuls ressorts. C'est cette ouverture proposée par M. Calon et A. Rip que nous recherchons dans le concept de dispositif socio-technique autant que l'opérationnalisation du concept.

3. Dispositif, concept clef en Sciences de l'information et de la communication ?

La question des dispositifs informationnels ou des dispositifs communicationnels est évoquée depuis longtemps dans le champ des SIC. Cette disjonction, selon la nature informationnelle et communicationnelle des dispositifs, est toujours opérante dans le champ des SIC. Elle donne une lecture asymétrique et orientée du dispositif privilégiant une approche scientifique, un type de ressources ou de fonctionnalités, informatives ou communicationnelles. Nous nous positionnons dans une perspective symétrique, attribuant aux notions d'information et de communication le sens de deux processus complémentaires lorsqu'ils se réalisent dans une situation.

Une manière de fixer cette symétrie consiste à composer les deux termes dans une expression bloquée, telle qu'information-communication. On trouve trace de la construction adjectivale *info-communicationnel* en 2002, dans un hors-série de la revue *Communication& Organisation* (HS. N°2) restituant les actes des débats de la journée du 23 octobre 1998 organisée en hommage à Robert Escarpit. Prenant la parole à l'occasion de la première table ronde portant sur la question : « *Quelles perspectives pour les Sciences de l'information et de la communication ?* », Bernard Miège suggère « […] *de mettre en œuvre une approche communicationnelle, ou on pourrait dire info-communicationnelle.* », comme réponse aux approximations et à l'accaparement technocentriste du discours sur les TIC (Hotier, 2002).

L'affirmation unitaire, qui se dégage de la juxtaposition connectée des deux termes information et communication, est avant tout et dans le contexte de son énonciation, la manifestation d'un projet disciplinaire. Cette forme condensée traduit une étape franchie dans l'affirmation d'une discipline et d'un projet scientifique unitaire.

La caractérisation scientifique d'un dispositif info-communicationnel apparaît à partir de 2006 dans des publications de chercheurs du LERASS (thèse de C. Gardies) qui conduiront à la publication de l'ouvrage collectif « *dispositifs info-communicationnels : questions de médiations documentaires* » sous la direction de Viviane Couzinet (Couzinet, 2009). Citant Jean Meyriat, V. Couzinet souligne que l'étude des dispositifs info-communicationnels est l'occasion d'un positionnement entre deux « *objets solidaires, mais distincts, de connaissances scientifiques* » (Meyriat, 1983). Organisé autour des trois finalités que sont la formation, la recherche et la culture, l'ambition de l'ouvrage est d'apporter des éléments de compréhension d'une classe de dispositifs étudiés dans La perspective des Sciences de l'information-documentations, selon la définition de Jean-Paul Metzger (Metzger, 2013). Il ne s'agit plus dans cet ouvrage de réaliser un instantané de l'utilisation du terme dans le champ d'étude mais de le rendre opérant en tant que concept.

L'apport de la notion de dispositif réside en premier lieu, dans le positionnement médian entre les deux objets information et communication. V. Couzinet, citant encore J. Meyriat évoque « *des objets*

solidaires, mais distincts, de connaissances scientifiques » (Meyriat, 1983) pour traduire la mise en tension du couple information communication (Couzinet, 2009).

Dans un second lieu, cette notion introduit un degré d'abstraction plus fort sur le fait technique. La distance plus importante aux TIC rééquilibre les discours qui se détachent des problématiques instrumentales et individuelles et donne une perspective plus globale, à même de restituer les enjeux.

Il y a dix ans, Bernard Miège soulignait que « *l'emploi de cette dernière notion [dispositif] est encore quelque peu prématuré : elle implique une complémentarité et une stabilité entre les éléments composant le dispositif, une articulation entre des outils et des contenus, et des usages bien spécifiés, in situ et à distance, sinon de façon ubiquitaire.* » (Miège, 2007, p48). Cet avis s'il est réservé, souligne néanmoins l'intérêt pour cette notion dans la discipline.

Dix années plus tard, nous pouvons constater que le terme dispositif est mobilisé de multiples façons dans la discipline, ce qui traduit si ce n'est l'émergence d'un concept clé, du moins la nécessité d'articuler le rapport entre les produits de la technique et leurs inscriptions sociales. Ainsi, nos collègues du laboratoire I3M de Toulon ont-ils forgé la notion de *DIspositif Socio-technique d'Information-Communication* (DISTIC) comme objet de recherche fédérateur d'un programme scientifique transversal[80].

Si l'on s'intéresse aux publications scientifiques, nous pouvons constater que le concept de dispositif fait l'objet d'un questionnement épistémologique récurrent dans les SIC comme l'illustre la publication de l'ouvrage collectif *Les dispositifs d'information et de communication – concept, usages et objets* produit par les chercheurs du Centre de Recherche sur les Médiations (CREM) sous la direction de Violaine Appel, Hélène Boulanger et Luc Massou aux éditions de Boeck (Appel, Boulanger, Massou, 2010).

Le contenu de cet ouvrage, comme celui antérieur du N° 25 de la revue *Hermès* publiée en 1999, permet de faire un état des lieux critique du concept. Comme V. Appel, H. Boulanger et L. Massou le soulignent, le concept de dispositif, qu'ils qualifient d'entre deux, pourrait contribuer à ne pas replier les questions de recherche qui animent les SIC sur des spécificités et des sous-domaines disciplinaires de plus en plus resserrés et cloisonnés (Appel, Boulanger, Massou, 2010).

Ces différentes réflexions associant le concept de dispositif et l'évolution de la discipline, nous ont amené à rechercher dans l'analyse de l'historique des publications de thèses les éléments d'une réflexion sur cette notion en SIC.

Depuis 2011, la base de données Theses.fr référence l'intégralité des thèses en préparation sur le territoire national. Par reversements successifs (dont les notices du catalogue du Sudoc) cette base est devenue exhaustive et reflète l'ensemble de la production des thèses depuis 1985. Cet outil bibliographique offre une vue unique sur la production scientifique de ces trente dernières années.

---

[80] http://i3m.univ-tln.fr/IMG/pdf/dispositifs-sociaux-tech-info-com-i3m.pdf consulté le 12 oct. 2015. Voir également le séminaire *Usages des dispositifs sociotechniques numériques* organisés par les laboratoires ELLIADD et OUN (Université Franche-Comté), ISCC (CNRS), I3M (Université Toulon).

Afin de mieux cerner la manière dont le concept de dispositif est utilisé dans les Sciences de l'information et de la communication, nous nous sommes livré à une analyse des données bibliographiques des thèses soutenues en SIC depuis 1985 et déposées sur le site www.theses.fr géré par l'Abes; (Agence bibliographique de l'enseignement supérieur).

D'autres chercheurs ont, avant nous, emprunté cette voie d'analyse de la discipline des SIC au travers des données bibliographiques des thèses (Polity, Rostaing, 1997), (Gallezot, Boutin, Dumas, 2006). Les premiers ont proposé une cartographie des thématiques de thèses soutenues durant 20 ans. Les seconds ont travaillé l'articulation disciplinaire information *vs* communication par l'analyse des sous-domaines identifiés à partir de l'indexation RAMEAU.

Conformément aux conditions générales d'utilisation (CGU)[81], l'interface proposée par l'Abes;[82] permet d'enregistrer les résultats d'une requête et de les exploiter. Cependant, la vue proposée du catalogue des thèses est réduite aux principaux champs identifiant les travaux et ne contient pas les informations bibliographiques secondaires : mots clés et résumés. L'accès à la notice complète d'une thèse nécessite une requête spécifique portant sur son identifiant. Pour recueillir un corpus de résumés ou de mots clés, il faut s'adresser à l'institution ou mettre en œuvre les techniques de la fouille du Web. Nous utilisons les outils de *scraping*[83] développés dans la plateforme d'analyse MEDIASWELL[84] pour automatiser cette extraction. L'intérêt est également d'illustrer le fonctionnement et les méthodes de mise en œuvre de ces outils.

Enfin cette étude met en œuvre des techniques d'analyse textuelle et en particulier l'outil IRaMuTeQ[85] développé au sein du LERASS[86] par Pierre Ratinaud. L'analyse textuelle n'est pas sans difficultés et demande beaucoup de prudence et de modestie dans l'interprétation des résultats.

3.1. Données d'analyse

Dans le cas présent, une partie des données d'analyse est obtenue à partir de l'enregistrement du résultat d'une requête au moteur de recherche de theses.fr[87]. La plateforme MEDIASWELL (cf. chapitre 4 et son complément) est utilisée pour récolter les résumés et les mots clés à partir des pages de présentation des données bibliographiques de chaque thèse identifiée par la requête précédente.

3.1.1. Méthode

La méthode employée comporte plusieurs étapes :

---

[81] http://www.theses.fr/conditions.html

[82] http://documentation.abes.fr/aidethesesfr/accueil/ch03.html

[83] Voir Ferrara & Al. Pour une revue sur les méthodes (Ferrara & Al, 2014)

[84] Voir les références personnelles à ce sujet.

[85] Disponible sous licence GNU GPL http://www.iramuteq.org/

[86] http://www.lerass.com/

[87] Nous désignerons par *données nominales* ce premier jeu de données

a) Établir une liste d'URL germes (*seed*) définissant une collection de référence, c'est-à-dire une présélection étendue de pages HTML du Web dont l'analyse des contenus établira ultérieurement la pertinence pour la constitution d'une collection ;

b) Constituer une base de données par la collecte et l'enregistrement de la représentation HTML calculée de chacune des pages Web associées aux URL précédentes. Ce calcul se fait dans le moteur de restitution Gecko (FireFox). Ce sont les technologies (Ajax) de la publication dynamique qui imposent cette interprétation par le cœur d'un navigateur pour assurer la complétude des informations des pages du Web. L'enregistrement de ces représentations HTML internes complètes construit une collection horodatée d'informations capturées dans un intervalle de temps proche (quelques heures). L'unité temporelle est une garantie d'homogénéité de la collection. Elle constitue une vue instantanée du Web pour les contenus ciblés (synchronicité). Par ailleurs, l'enregistrement intégral des pages permet l'analyse hors ligne des contenus.

c) Extraire à partir des pages enregistrées, les blocs structurels HTML comportant les informations pertinentes et les stocker dans une collection de fragments analysables (*i.e.* : résumé, mots clés, etc.). Il s'agit d'une étape intermédiaire qui n'a véritablement de sens que dans le cas de traitements massifs de pages dont la structure est peu maîtrisée *a priori*. Elle permet de vérifier et de filtrer les enregistrements de pages en fonction du degré de leur complétude. Suivant ce degré, l'étape précédente peut être réactivée ou la page invalidée. Enregistrer un éclaté de la structure de la page, suivant une maille plus grossière, en sélectionnant des nœuds HTML plus proches de la racine, permet de rattraper des variations structurelles qui seront analysées dans un second temps.

d) Extraire les données brutes (*row data*) à partir des blocs segmentés. L'extraction est effectuée par un automate dont le paramétrage est réalisé au moyen d'une grammaire. Pour chacun des blocs informationnels, cette grammaire permet de décrire un ensemble de règles alternatives explicitant les différents chemins (voire un seul) conduisant à l'extraction des données brutes[88]. Donner un ensemble de règles permet d'adapter l'extraction à la variabilité des pages. En général quelques cas suffisent pour couvrir l'intégralité des variations rencontrées.

e) Raffiner les données brutes pour construire les données de référence à partir desquelles s'organiseront les données d'analyses et les exports. Cette étape permet de lisser les représentations des données lorsque celles-ci sont susceptibles d'expressions alternatives (par exemple les dates, etc.).

[88] Le formalisme de référence pour décrire la grammaire est le formalisme JSON. Les expressions grammaticales peuvent contenir des expressions régulières. Par exemple : {"prof": '#<div id="experience-[\d]*-view>'} signifie que la localisation de la donnée brute associée à l'élément structurel "prof" est associée à une balise de subdivision ( ) du bloc et que cette balise a pour identifiant ('id') une valeur littérale dont le préfixe contient "experience-" suivi d'une valeur numérique suivi du suffixe "-view" (ex : "experience-1234-view" est une expression valide.)

Les étapes b), c), d) conduisent à la création de tables d'enregistrements distinctes en base de données. Les tables intermédiaires sont des facilités pour assurer l'enchaînement ou la reprise des traitements pour une même collection.

3.1.2 Collecte des données

Dans le cas de l'analyse des notices de la base Theses.fr, il n'y a pas de variabilité structurelle notable des pages. En effet, celles-ci sont produites à la volée par le serveur et leur structure est maîtrisée par les concepteurs du site. Seule la saisie en texte libre, durant la phase de description de la thèse, introduit des variations (orthographiques, typographiques, lexicales) affectant la désignation des entités référentielles : directeur de thèse, université ou discipline.

Collecte des données nominales étape a)

Elle est réalisée à partir d'un export des données[89] associées à la recherche avancée portant sur la discipline : "information (AND) communication" et concernant les thèses soutenues. On obtient (en date du 23/10/2015) 1849 thèses recensées dans le catalogue (sur un total de 2077 inscriptions). Parmi ces thèses, toutes ne correspondent pas au champ disciplinaire des SIC. Certaines traduisent des parcours individuels complexes (plusieurs réinscriptions, etc.). Il est donc nécessaire de clarifier les données, au plus tôt, à partir des informations contenues dans le fichier résultat.

L'analyse de ce fichier est réalisée à l'aide du logiciel OpenRefine[90] qui permet de normaliser rapidement les formes lexicales. La normalisation a porté sur l'expression disciplinaire qui, de manière surprenante, connaît beaucoup de variations autour des termes "information" et "communication", y compris à l'intérieur des SIC. Elle a porté également sur la codification de l'établissement et du lieu.

[89] 3 exports fractionnés sont nécessaires, les résultats délivrés par theses.fr ne dépassant pas 1000 réponses consécutives.

[90] Logiciel open source pour travailler sur des données désorganisées (*messy data*). OpenRefine fait suite à GoogleRefine : http://openrefine.org/

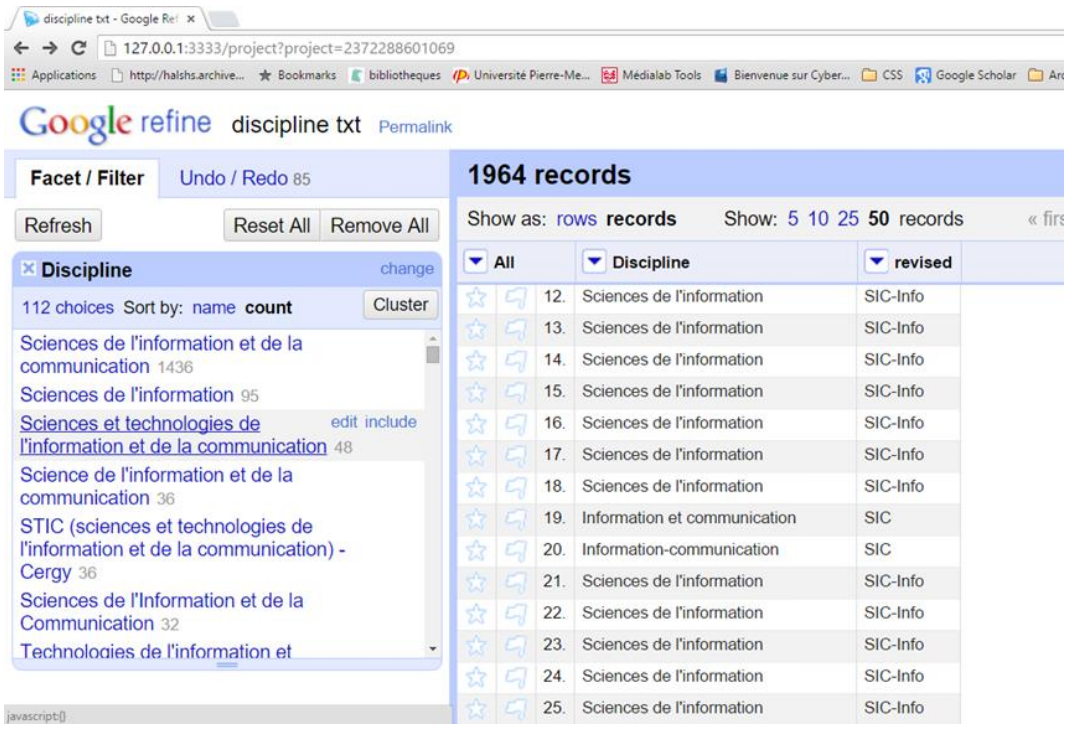

Fig1. Utilisation d'OpenRefine
pour traiter de la variabilité de l'expression.

À l'issue de cette étape, il ne reste que 1625 thèses s'inscrivant dans le champ disciplinaire des SIC[91]. La liste des URL de pages candidates est engendrée automatiquement à partir du fichier précédent[92].

Cette liste d'URL guide l'étape b) qui est réalisée dans MEDIASWELL par un module spécialisé dans la navigation Web.

Scraping des pages : étape c) et d)

L'étape c) n'est justifiée que pour vérifier que l'enregistrement de la page s'est bien passé (quelques pages ont nécessité d'être reprises). Les étapes de *scraping* c) et d) sont réalisées dans la foulée. L'étape c) est réalisée par un module générique alors que l'étape d) nécessite une adaptation légère pour cibler les contenus.

[91] Les travaux se rapportant aux STIC ou aux Sciences de l'éducation sont éliminés. Quelques travaux revendiquant la bi-appartenance ont été conservés.

[92] L'URL d'une page est obtenue en ajoutant le code interne de l'enregistrement de thèse à l'URL du site.

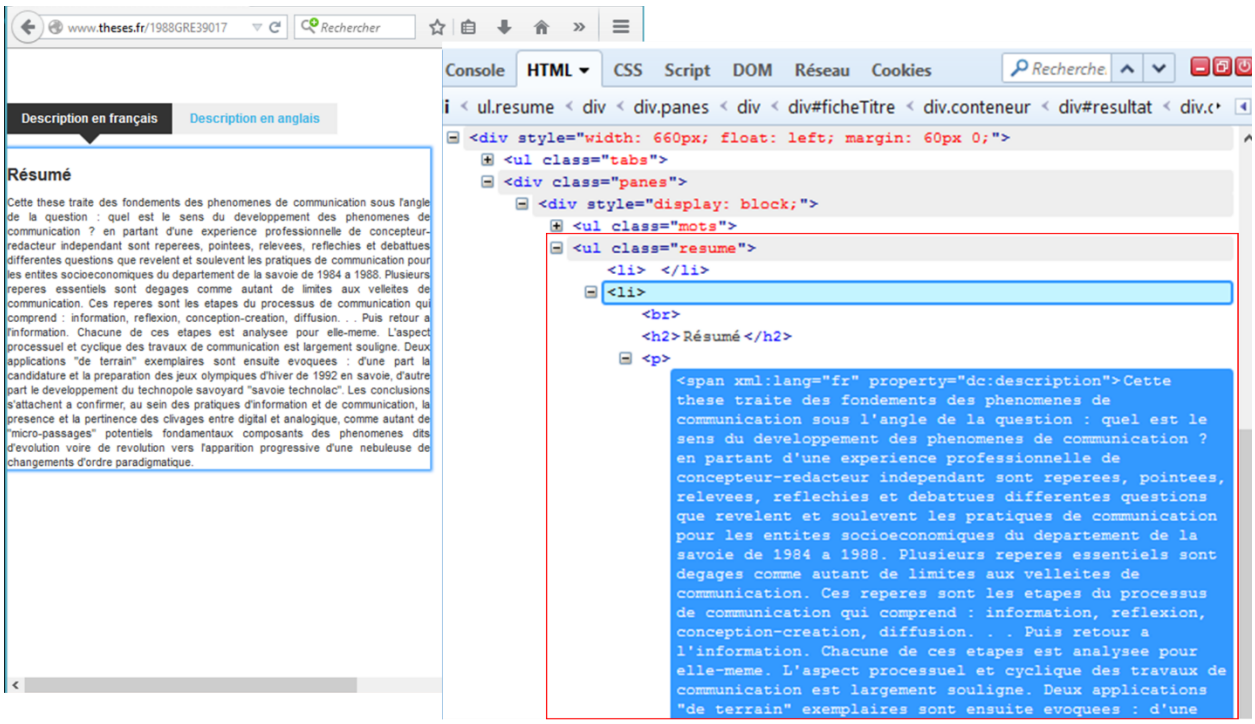

Fig2. Vue réalisée par le moteur gecko
pour un enregistrement de theses.fr. (plugin Firebug)

Les représentations HTML extraites du moteur graphique gecko ont une structure logique claire (cf. Fig2) : chaque élément de la notice bibliographique correspond à une subdivision identifiée par une classe unique. Ainsi le résumé est identifiable à l'aide de l'expression grammaticale : {"résumé": '<ul class="resume">'} associant la classe "resume" à la balise d'alinéa de liste (<ul>). La grammaire mise en place pour l'extraction des données brutes traduit les objectifs de l'analyse qui porte sur le résumé et les mots clefs. Les autres informations présentes sur la page sont redondantes avec celles obtenues à l'étape a).

À l'issue des étapes c) et d), les données normalisées sous OpenRefine sont ajoutées aux données textuelles (titre, résumé, mots-clefs) extraites des pages Web pour constituer les données de référence. Deux indicateurs numériques correspondant respectivement, au nombre d'occurrence du terme "dispositif" dans le titre ou dans le résumé, sont associés à chacun des enregistrements.

L'étape e) n'est pas nécessaire dans cet exemple, les textes des résumés ou les mots clés (données brutes), n'appelant pas de retraitements spécifiques.

En définitive, il ne reste que 1130 enregistrements de thèses en SIC, dont le titre et le résumé sont définis et à partir desquels nous pouvons envisager une analyse fondée sur des statistiques textuelles. Parmi elles, 229 thèses comportent une référence au terme dispositif, dont 40 dans l'intitulé (titre).

3.2. Analyse

3.2.1. Analyse statistique

L'analyse statistique est réalisée à partir du fichier consolidé des réponses (étape a).

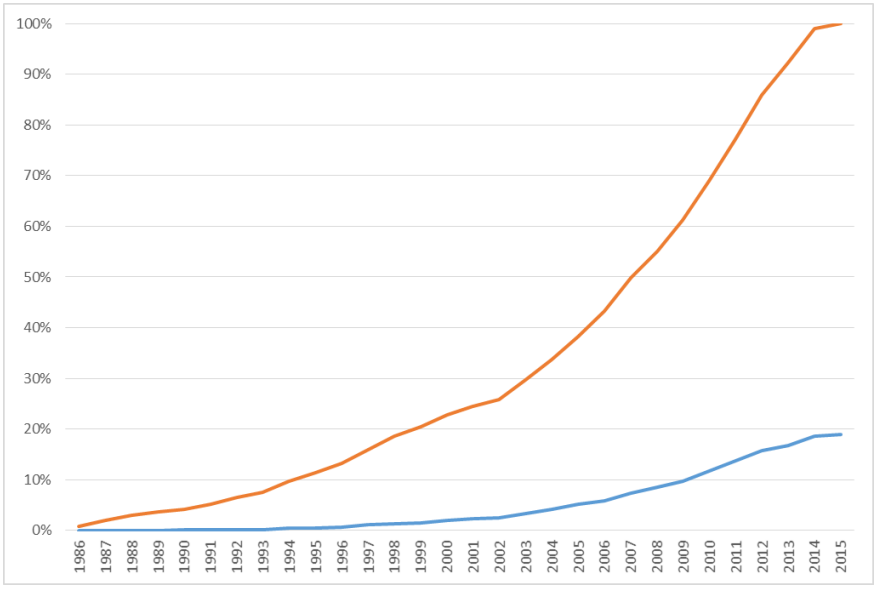

Fig3. Évolution annuelle cumulée du nombre de thèses soutenues

(orange) exprimé en pourcentage du total et évaluation annuelle cumulée du nombre de résumés mentionnant un dispositif (titre ou résumé) rapporté au nombre de thèses cumulées (bleu).

La figure précédente (Fig3). Montre une augmentation du nombre de thèses soutenues dont la croissance se rapporte à une loi de puissance. Cette progression s'accompagne d'une part croissante du nombre de thèses consacrées à la notion de dispositif que l'on peut approcher par une loi de type polynomiale (ordre 2). Cette évolution du nombre de thèses consacrant la notion de dispositif rend compte d'un usage plus marqué en Sciences de l'information communication dès la fin des années 1990. L'apport de l'Internet et du Web n'est pas aussi significatif qu'il y parait : seulement 14 résumés (1%) comportent une mention de l'un des termes "Web" ou "Internet", dans le titre ou le résumé. On arrive à moins d'une cinquantaine d'éléments en ouvrant à des notions tels que "réseau", "net" ou "service" (4%). Bien que l'on recouvre ainsi près de 25% du nombre de thèses ayant une occurrence du terme dispositif (résumé ou titre), l'explication est à rechercher de manière plus large et pas seulement dans le champ technique.

L'analyse de la répartition spatiale des thèses soutenues, met en évidence la très forte présence de la notion de "dispositif" dans des travaux associés au Centre Norbert Elias de l'université d'Avignon, dont les membres ou membres associés étaient directeurs de thèse principaux pour 23 thèses soutenues : J. Davallon (9), E. Ethis (5), Y. Jeanneret (3), D. Jacobi (2), B. Dufrêne (2), H. Gottesdiener (2).

Localisation	Présent	Absent	total	% Présent
Paris	78	409	487	16%
Lyon	16	92	108	15%
Bordeaux	6	63	69	9%
Grenoble	10	47	57	18%
AixMarseille	10	31	41	24%
Rennes	13	27	40	33%
Avignon	23	16	39	59%
Toulon	7	31	38	18%
Montpellier	8	28	36	22%
Metz	10	25	35	29%

Fig4. Nombre de thèses soutenues contenant ou non le terme dispositif

parmi les 10 lieux ayant le plus de thèses enregistrées.

Sous l'impulsion de ces chercheurs, les travaux du centre Norbert Elias s'inscrivent dans une tradition à forte dominante culturelle et artistique (muséale et patrimoniale). Il est probable que cette filiation contribue au développement du concept dans ce champ d'actions.

3.2.2. Analyse textuelle

À partir de la base de données réalisée à cet effet, il est possible de produire deux corpus (extractions) associés aux thèses soutenues : l'un porte sur les titres, l'autre sur les résumés. Afin de conduire une analyse sous IRaMuTeQ, deux fichiers distincts sont engendrés. Ces fichiers contiennent l'ensemble des textes (intitulé ou résumé) sélectionnés, complétés de variables indicatives (lieu d'inscription, nombre d'occurrences du terme cible, etc.).

Pour simplifier l'analyse des séquences d'indexation (mots clefs), celles-ci sont assimilées à des textes, ce qui permet de réaliser une analyse sous IRaMuTeQ.

Analyse des titres de thèses.

Parmi les 40 descriptifs de thèse dont le titre mentionne explicitement le terme "dispositif", 2 n'ont aucune mention du terme dans le résumé. Nous privilégierons dans l'analyse des titres, les 38 enregistrements couvrant titre et résumé. L'hypothèse sous-jacente est qu'un terme qui apparaît à la fois dans le titre et le résumé est par nature non substituable et devrait correspondre à une notion.

L'analyse des similitudes (ADS) consiste à étudier les proximités statistiques au sein d'un graphe de relation. Dans le cadre de l'analyse textuelle, cette méthode permet de mettre en évidence les couples de termes fortement co-occurrents dans les énoncés (Marchand, Ratinaud, 2012).

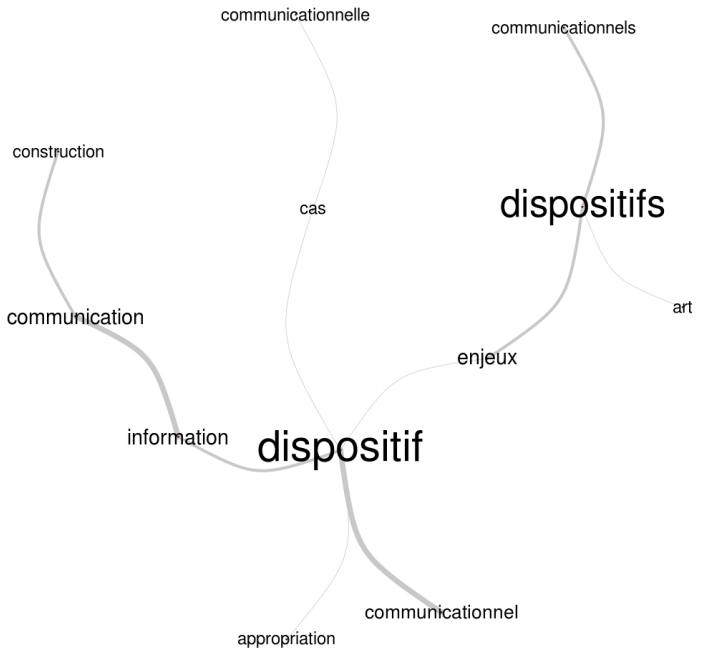

Fig5. Graphe de similitude des titres de thèse
mentionnant "dispositif" (sans lemmatisation).

Ce graphe montre en premier lieu la dominante communicationnelle associée à la notion de dispositif. Dans le contexte d'emploi au pluriel, le terme d'information parait se dissoudre dans le

contexte. Il est en revanche davantage associé à la définition d'un dispositif dans le cas singulier. La différence singulier/pluriel semble importante et traduire des usages distincts. Cette distinction se retrouve également dans l'indexation des documents où les deux formes sont concernées. Elle semble traduire deux approches différentes du dispositif en tant que classe, l'une étant définie en intension (singulier) alors que l'autre le serait en extension (pluriel).

L'analyse des 207 résumés comportant au moins une mention du terme dispositif permet d'aborder plus finement les contextes d'usage.

Au préalable, on procède à une évaluation de la variabilité des textes résumés en comptant le nombre de mots au regard de deux indices de la diversité du vocabulaire : le nombre de formes[93] présentes (bleu) et le nombre de formes uniques (orange).

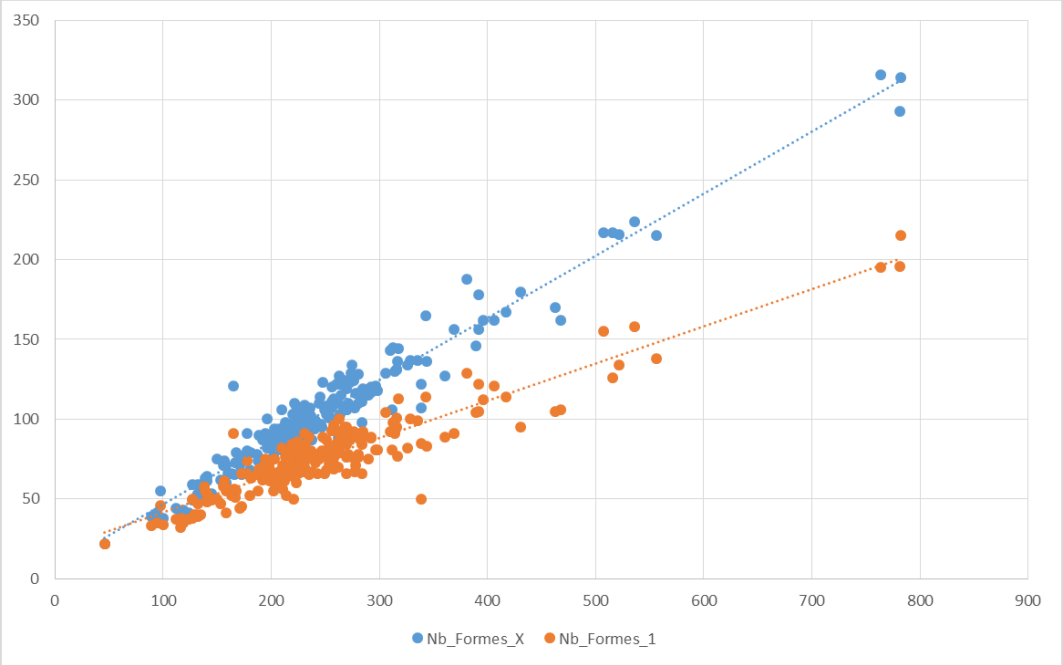

Fig6. Droites traduisant la variation linéaire
du nombre de forme en fonction de la taille du texte.

Ce graphe présente la forme habituelle de l'expression de la diversité du vocabulaire dans un texte. Pour mieux évaluer la qualité de l'AFC qui suit nous avons éliminé le premier et le dernier décile constituant un second corpus de validation plus concentré.

La classification hiérarchique descendante pratiquée sur ces deux corpus, selon la méthode de Reinert (double sur RST[94]) permet de réaliser des regroupements de segments de textes en fonction de la fréquence de leurs cooccurrences (Reinert, 1983), (Ratinaud, Marchand, 2012).

La comparaison des deux résultats produits avec chacun des deux corpus permet d'éprouver la stabilité de l'AFC. Ce qui est bien le cas présent.

---

[93] Nous avons travaillé à partir de l'analyse en TAL (module développé avec le LIDILEM) sur les catégories de TreeTagger : NOM, NAM, ADJ, ADV, VERB.

[94] Méthode SVD : irlba, dictionnaire : indexation, Taille rst1 : 12, rst2 : 14,
nombre de classes terminales phase 1 : 10

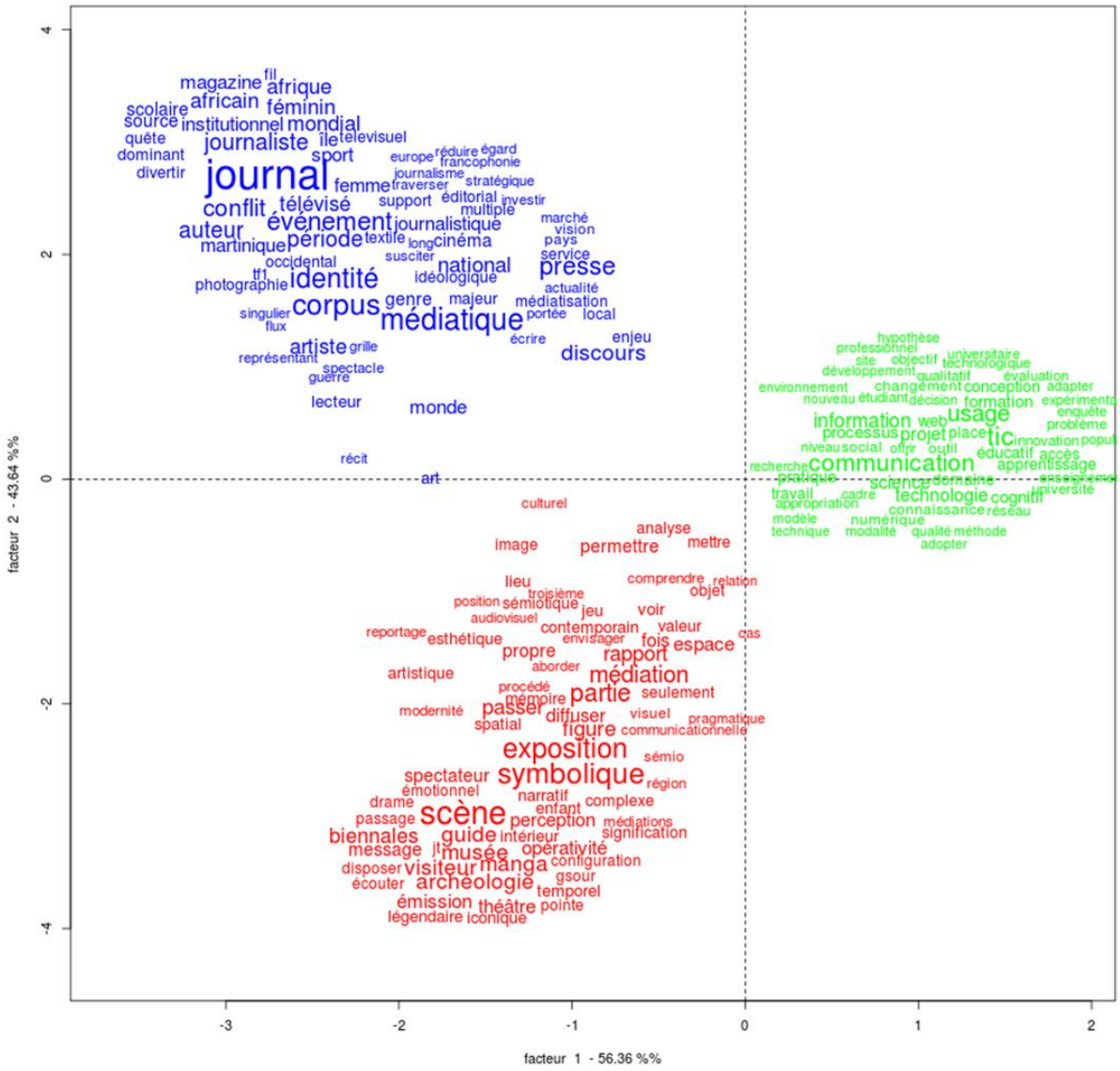

Fig7. Analyse Factorielle des Correspondances (AFC)
associée à la classification des résumés

Cette analyse fait clairement apparaître 3 classes terminologiques[95] dans lesquelles est convoquée la notion de dispositif :

- C1 : (rouge) : en rapport avec les logiques de médiation, d'exposition et de mise en scène[96].
- C2 : (vert) : en rapport avec les TIC, leur développement et leurs usages.
- C3 : (bleu) : en rapport avec les journaux (et médias), leurs rôles et leurs productions.

Ce regroupement dans un nombre de classes minimal est nécessaire pour utiliser ce résultat. Ainsi, les 3 classes issues de l'analyse sous IRaMuTeQ sont codées et reportées sur le descriptif des enregistrements dans la base de données. Cet enrichissement permet de relancer une analyse sur

[95] Permettant la catégorisation de 201 des 207 résumés soit 97% du corpus, ce qui est une excellente couverture pour une analyse factorielle.

[96] On retrouve notamment dans la classe 1 la très forte représentativité des travaux réalisés au centre Norbert Elias et, également dans une moindre mesure, à Dijon et à Tours.

les textes complets et non plus seulement sur les segments représentatifs de ceux-ci. Augmenter le nombre de classes conduirait, dans le cas présent, à la constitution d'échantillons statistiques devenus trop petits.

En opérant sur une sélection suivant les différentes catégories des résumés, on peut s'appuyer sur une analyse des similitudes dans les 3 contextes d'usages du terme dispositif et interpréter la nature du clustering.

Le premier axe factoriel oppose la classe 2 aux deux autres. Plusieurs interprétations sont plausibles. Une première opposition semble porter sur la différence entre d'une part, les TIC considérées comme objet (instrument) et participant parmi d'autres à un dispositif, et d'autre part, des dispositifs constitués (scènes, journaux, etc.). L'institutionnalisation parait également un élément de cette opposition.

Le second axe factoriel souligne l'opposition entre les classes 1 et 3. Une interprétation possible porte sur la fonction de médiation (culturelle, artistique) mise en avant dans la classe 1 qui donne aux dispositifs de représentation un sens par eux-mêmes. À l'opposé, le journal/média (institution) est envisagé davantage comme un dispositif dont on étudie la construction, notamment au travers des publications (corpus).

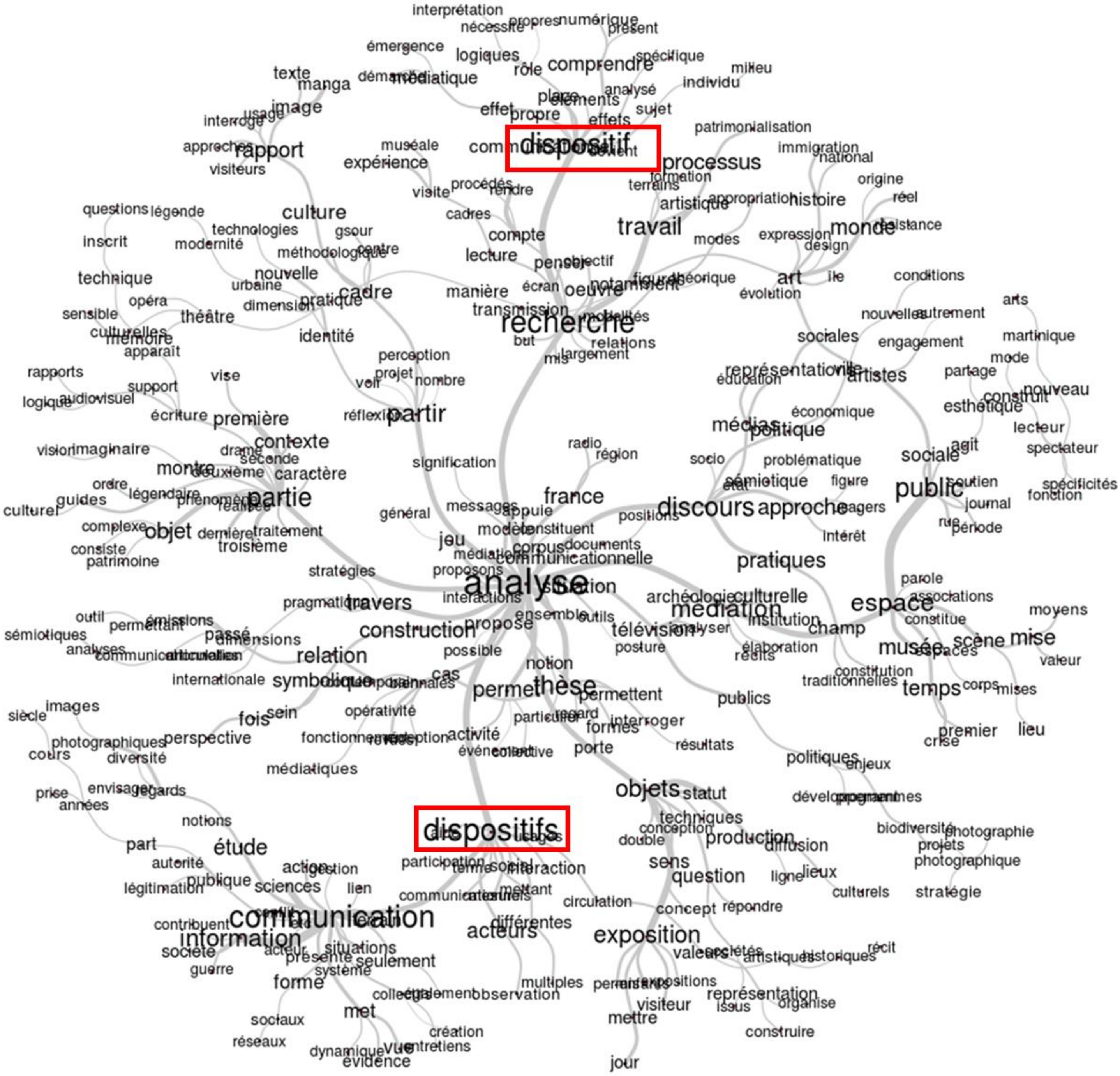

Fig8. Analyse des similitudes (cooccurrences)

dans la classe 1 (Freq. > 5) 52 résumés, 3522 formes, 2150 hapax

(12.79 % des occurrences – 61.04% des formes)

La forme singulier est associée (fortement) à l'activité et aux questions de recherche alors que la forme plurielle porte les enjeux d'analyse info-communicationnels dans le domaine concerné.

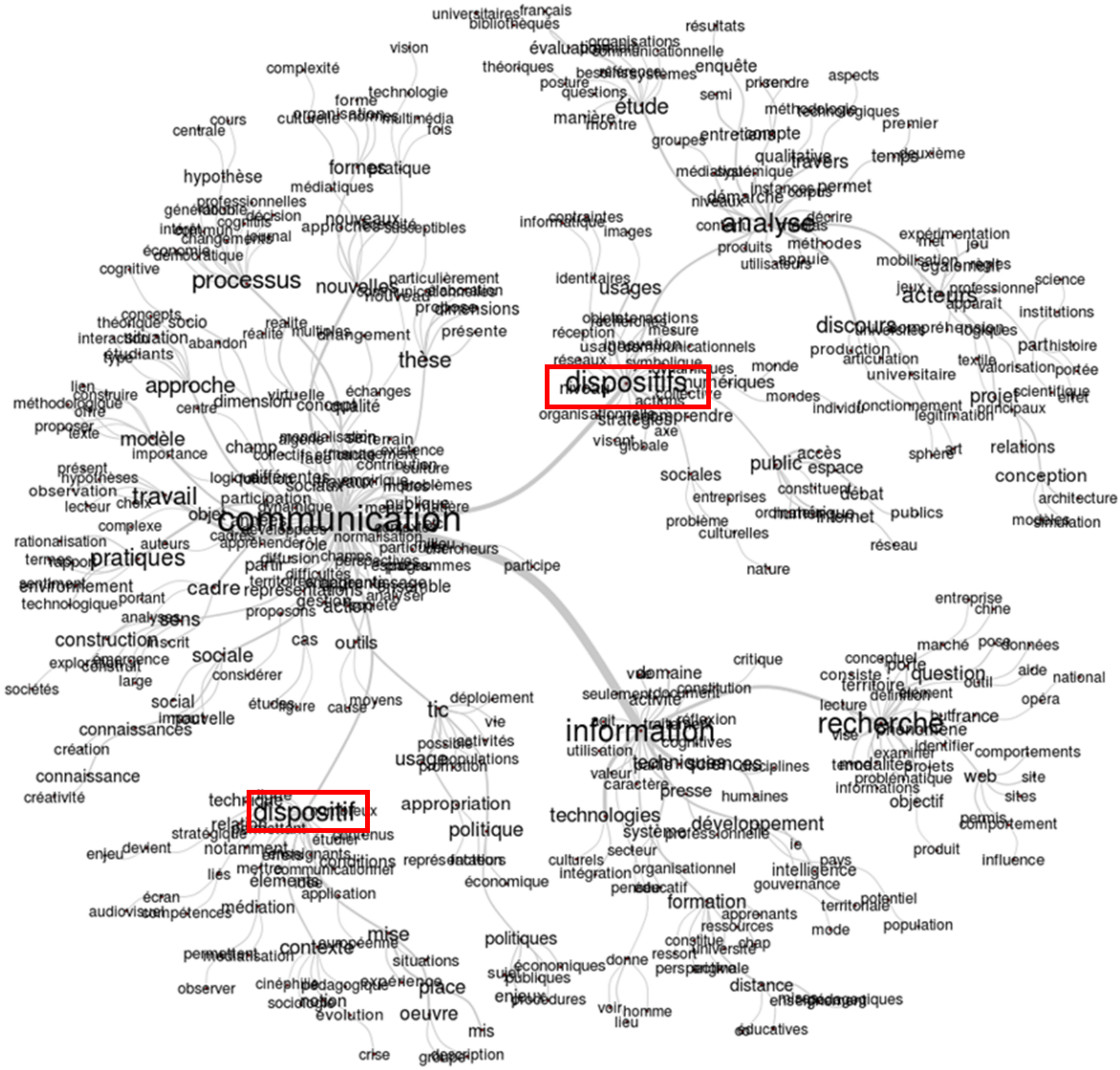

Fig9. Analyse des similitudes (cooccurrences)
dans la classe 2 (Freq. > 5) 105 résumés, 4865 formes,
2789 hapax (10.03 % des occurrences – 57.33 % des formes)

Les deux formes singulier et pluriel sont articulées au terme "communication". Cette symétrie rend difficile la distinction entre objet et méthode. Il semble toutefois que la forme plurielle renvoie davantage à l'objet plutôt qu'à la méthode.

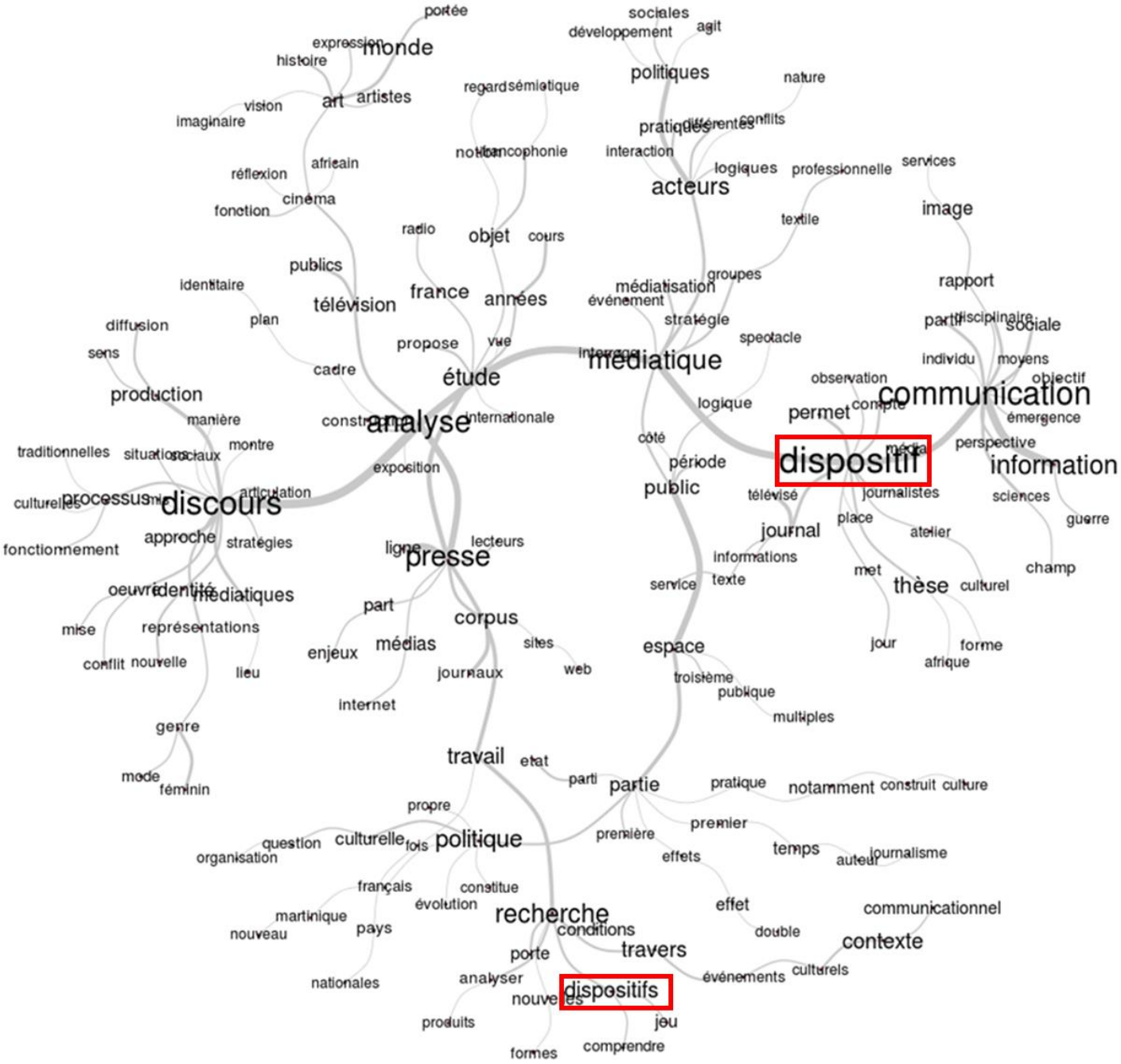

Fig10. Analyse des similitudes (cooccurrences)
dans la classe 3 (Freq. > 5) 44 résumés, 3061 formes, 1951 hapax
(15.60 % des occurrences – 63.74% des formes)

Dans ce graphe beaucoup moins dense que les autres, la forme singulier est dominante. Elle est clairement associée à l'objet d'étude (journal) qui est posé en dispositif.

3.2.3. Analyse des mots clés

Les mots clés sont le résultat d'une indexation libre des thèses réalisées par leurs auteurs (*folksonomie*). Ils sont portés à la fois par une intuition et un savoir-faire documentaire autant que par l'ambition d'un référencement et d'une diffusion large des travaux. Dans le cadre de l'étude menée, l'analyse des formes de l'indexation personnelle ne présente pas d'intérêt. Elle nous permet

de dégager du sens globalement, dans l'analyse synthétique que l'indexation construit des sous-domaines et des champs de recherche.

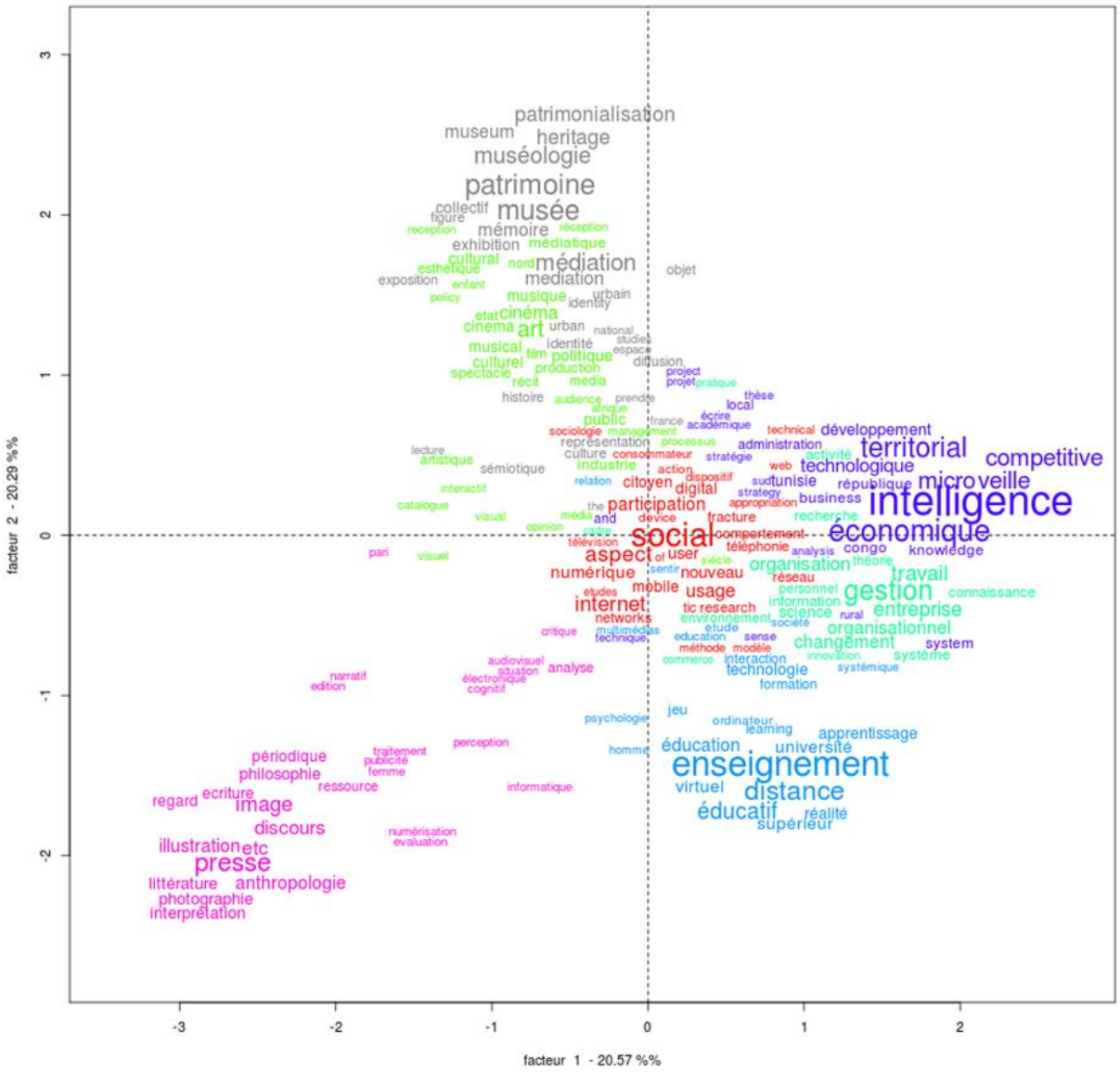

Fig11. Analyse Factorielle des Correspondances (AFC)
associée à la classification des séquences d'indexation des thèses dont le résumé comporte une mention du terme dispositif.

La classification mise en œuvre est très fine ; en est exclue une grande partie des documents, ce qui n'était pas le cas Figure 6. On retrouve néanmoins des oppositions identiques à celle de l'AFC des résumés, pour les classes 2, 6 et 7. Le premier axe factoriel nous semble traduire la même opposition associée à la référence technique liée aux Sciences de l'information et aux TIC.

La classification est plus fine qu'elle ne l'a été pour les résumés. Elle fait ressortir 7 grandes classes thématiques de travaux mobilisant le terme de dispositif. Le dendrogramme associé illustre la nature des classes et leur cohérence avec les 3 classes retenues pour les résumés.

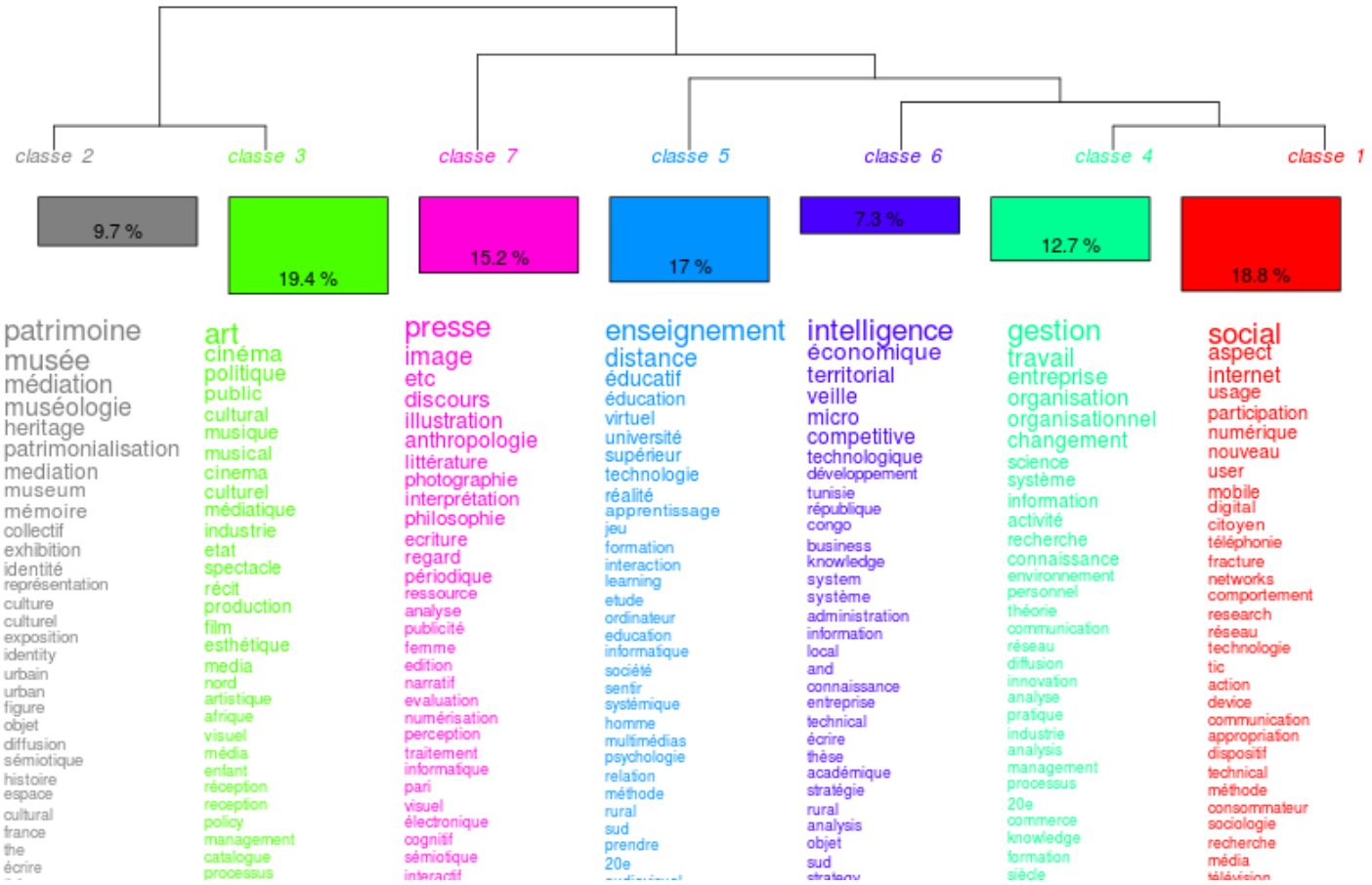

Fig12. Dendrogramme associé à l'analyse factorielle Fig10.

La structuration des classes que l'on obtient ici n'est pas exactement une déclinaison des 3 classes obtenues lors de l'analyse des résumés.

La cartographie thématique qui se dégage dans les 7 classes de thèses que nous retenons comme référence (Fig12), n'est pas éloignée dans ses contours du panorama dressé dans l'ouvrage collectif du CREM (Appel, Boulanger, Massou, 2010). On retrouve des grands sous-domaines et les thématiques fortes des SIC dans les secteurs de la culture et de l'art, du journalisme et des médias, des technologies éducatives, de l'information et de la représentation des connaissances, de la communication institutionnelle et de l'espace public.

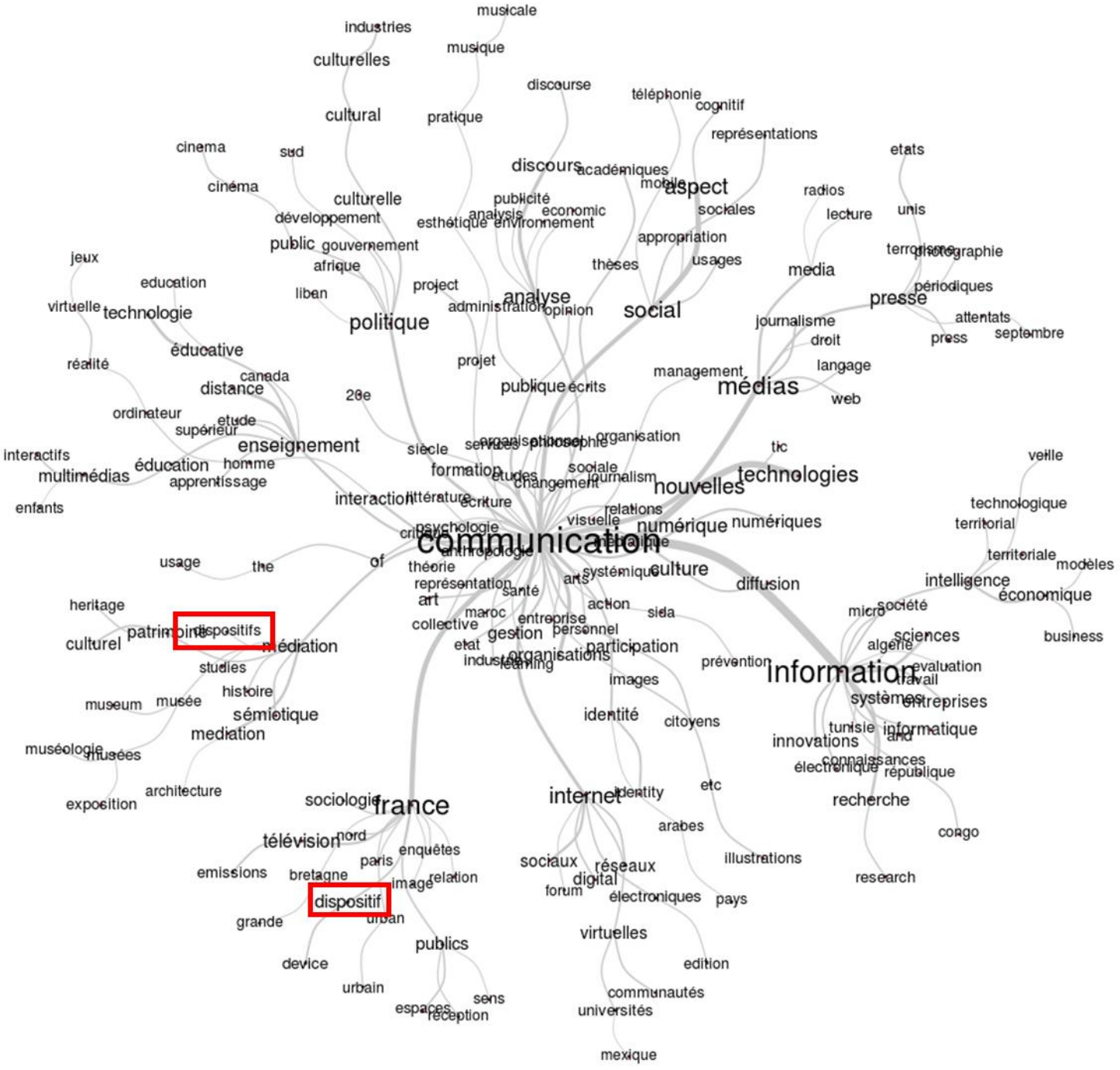

Fig13. Analyse des similitudes (cooccurrences)

des séquences d'indexation mots clefs (Freq. > 4) 224 séquences, 1553 formes, 925 hapax (17.48 % des occurrences – 59.56% des formes)

Cette analyse fait ressortir la distinction déjà relevée entre les deux emplois singulier et pluriel. On retrouve l'usage pluriel (cf. Fig7) de la classe 1 pour désigner l'objet d'étude et l'usage singulier (cf. Fig9) associés aux médias.

3.3. Résultats d'analyse

L'analyse statistique (Annexe A1 §2.1) apporte deux résultats. En premier lieu, l'augmentation du nombre de thèses soutenues annuellement en SIC traduit la vivacité du domaine et son expansion sur le territoire national. En second lieu, la croissance remarquablement forte du nombre d'occurrences du terme dispositif indique une inscription probablement plus forte du concept dans le domaine et les sous-domaines des SIC. Par ailleurs, si dans près de 25% des cas, la présence du terme dispositif est associée (cooccurrence) à celle d'un terme du lexique de l'Internet, la technicisation liée au développement du numérique, ne justifie pas à elle seule cette augmentation de la fréquence du terme dans le discours en SIC.

Enfin, l'analyse des intitulés de thèse, nous conduit à défendre l'hypothèse que le choix privilégié de la forme lexicale "dispositif" ou de "dispositifs" traduit un positionnement conceptuel différent. Cette hypothèse repose sur la nature documentaire du corpus. La notice enregistrée dans Theses.fr assure une fonction de communication institutionnelle et personnelle évidente. La construction du titre, le choix des mots clés, la proposition de résumé, sont établis par l'auteur de la thèse pour restituer de manière synthétique la nature de son travail, les enjeux s'y rapportant et la contribution scientifique et disciplinaire qu'il apporte. Les objets de recherche et les concepts clefs figurent nécessairement dans cette production sous une forme plus ou moins fixée témoignant de la maturité du concept et de sa force dans le domaine.

Par la suite, nous distinguerons le terme dispositif (lexique) des déclinaisons conceptuelles qui lui sont associés en leur donnant la forme de hashtag. Ainsi, #dispositif, #dispositifs traduisent deux propositions distinctes. La première est de nature holistique et l'unité dégagée mérite d'être questionnée. Elle peut renvoyer à une théorie dominante, ou à tout autre élément historique conduisant à unifier et lisser les différences. La seconde nous parait se rapprocher de la définition mathématique de classe, c'est-à-dire d'un regroupement dont il est possible de parler collectivement mais qui ne forme pas nécessairement un ensemble. On peut d'ailleurs relire la table des matières de l'ouvrage collectif dans ce sens et remarquer, que le chapitre 1 qui concerne la définition du concept, utilise une forme ambivalente de "dispositif(s)" et que le chapitre 9 utilise clairement la forme singulier, parlant du dispositif médiatique.

V. Appel et T. Heller, évoquent un sens du terme dispositif faisant référence aux processus identifiés à l'occasion d'une recherche suivant deux orientations analytiques possibles : « *la co-construction du social/technique et la co-construction de sens.* » (Appel, Heller, 2010). Selon ces auteurs, cette distinction repose sur deux filiations épistémologiques de la recherche en SIC, issues de la sociologie et des sciences du langage (*Ibid.*). On peut se demander si la distinction apparente entre #dispositifs et #dispositif ne fait pas écho à la distinction épistémologique précédente.

Si l'on considère que le dipôle information *vs* communication caractérise l'évocation disciplinaire dans les discours en SIC, alors l'inscription en périphérie de l'un ou l'autre des pôles traduit l'éloignement/proximité de la notion.

Le graphe des similitudes (Fig5) souligne la prédominance et la stabilité d'une lecture communicationnelle des deux termes. Dans le cas présent, c'est #dispositif qui semble s'imposer. L'analyse des similitudes appliquées sur les trois classes de résumés explicitées par l'analyse

factorielle (Fig7) met en évidence la différence des deux notions et leur mobilité dans l'ancrage disciplinaire suivant les sous-domaines :

- Classe C1 : dans le sous-domaine de l'art et de la culture (Fig8), #dispositifs domine les discours, #dispositif paraissant lié à l'activité empirique ou expérimentale (étude de cas) de la recherche. La notion de #dispositifs semble clairement posée. La très forte participation (59%) du nombre de thèses du centre Norbert Elias (Avignon) faisant référence à ce terme traduit également la volonté de fixer le concept dans le champ théorique ;
- Classe C2 : dans le sous-domaine des TIC et de leurs usages (Fig9), les deux termes sont attachés à la polarité communication. Avec réserve, la notion de #dispositifs nous semble devoir jouer un rôle voisin de celui jouer en C1. On peut probablement évoquer une instabilité chronique propre à une thématique en liaison avec l'innovation socio-technique, ne permettant pas de poser le concept ;
- Classe C3 : dans le sous-domaine médiatique (Fig10), c'est clairement la notion de #dispositif qui s'impose. La liaison avec les termes "médiatiques", journal (télévisé) témoigne de l'approche systémique privilégiée dans l'étude des médias[97]. Cette lecture est renforcée de la position marginale de #dispositifs. Les rôles respectifs que tiennent #dispositif et de #dispositifs dans l'organisation du champ lexical paraissent s'inverser par rapport à la classe C1 (voire aussi C2).

Cette différence entre C1 et C3 apparaît également dans le graphe des similitudes appliquées aux mots clefs (Fig13).

La distinction des deux occurrences #dispositifs et #dispositif dans la description de thèses en SIC permet de dégager de grandes tendances, que l'on peut résumer ainsi : l'unité d'un projet scientifique (C1), l'instabilité inhérente à l'innovation portée par les TIC (C2) et l'historicité de la notion dans le champ d'étude des médias (C3).

4. Opérationnalisation du concept

Si la notion de dispositif a pu apparaitre comme un mot valise, les volontés de rendre opérationnel le concept se sont multipliées ces dernières années et en font ressortir le potentiel heuristique. Le début des années 2000 a vu ressortir une complexité plus importante liée à la place croissante des réseaux informatiques dans la mise en relation des individus et des services. La capacité des TIC à s'agréger entre-elles n'en est pas la seule raison, la dimension technique se complétant désormais d'autres enjeux, symboliques et relationnels (Paquienséguy, 2007). Dès lors, il s'agit d'aller au-delà de la définition consensuelle d'agencement hétérogène d'objets techniques et d'individus interagissant. Opérationnaliser cette notion c'est avoir le projet de « *tenir ensemble les différentes dimensions qui instituent les pratiques sociales dans un contexte technique et symbolique hétérogène dont l'analyse*

[97] Voire à ce sujet la définition proposée par Y. Jeanneret de la notion de dispositif (Jeanneret, 2005).

des différentes parties ne peut se concevoir que dans leurs relations et leurs médiations. » (Monnoyer-Smith, 2016, p30)

Il nous parait nécessaire de poursuivre le travail d'opérationnalisation du concept, en le confrontant aux situations réelles et aux cadres d'analyses multiples que celles-ci engendrent. Dans le contexte du Web, Laurence Monnoyer-Smith identifie quatre dimensions selon lesquelles cet objectif peut s'envisager. La présentation qu'elle fait de ces dimensions ouvre des perspectives distinctes qui requièrent pour chacune d'elles un appareillage méthodologique qui reste à préciser (*Ibid.*, p24).

L'analyse des traces numériques d'usage offrent des possibilités ou du moins des pistes favorables d'investigations qui nous paraissent convenir avec plusieurs de ces axes. Cela apparait de manière évidente avec les énoncés produits par les dispositifs info-communicationnels qui s'exportent sur des fils de publications et que l'on peut collecter. Grâce aux collections réalisées, on peut espérer faire ressortir les logiques d'une production assimilable à celle du dispositif et ainsi tracer, sinon contribuer à rendre apparentes, les *lignes de visibilité*[98] du dispositif. Les mises en relations explicitées de thématiques, de connaissances, etc. qui structurent les échanges dans les plateformes collaboratives (réseaux sociaux, etc.) peuvent à leur tour être le moyen du repérage et de l'analyse de *lignes de force(s)* voire des *lignes de fuite*. Bien évidemment, il est difficile d'affirmer comme réalisables, en toute généralité, de telles hypothèses largement tributaires des spécificités des dispositifs et des circonstances de leur fonctionnement. Cependant, ces axes nous permettent d'éclairer certains résultats encourageants que nous avons pu produire. En particulier, la mise en évidence de dispositifs liés à l'activisme politique se développant sur Twitter dans un contexte de campagne électoral illustre parfaitement cette possibilité (Ref.35).

Selon l'approche empirique que nous menons, l'opérationnalisation a un sens beaucoup plus contraint. Elle est associée d'une part, à la production de modèles calculables fondés sur des traces numériques engendrées par le dispositif en fonctionnement, et d'autre part, sur une capacité à interpréter les résultats produits par ces modèles.

De ce fait, il est nécessaire d'affiner les notions en privilégiant celles faisant ressortir des observables pertinents.

Une première suggestion nous est faite par la généralisation des mécanismes de traçage qui sont opérés par les services du Web. Nous soutenons que pour les plateformes contemporaines du Web, le traçage personnalisé est une manière de se constituer en dispositif. Le traçage est fondamentalement au service d'une connaissance qui inclue dans sa production les usagers. Les mécanismes d'implication de l'usager-abonné sont multiples, allant de l'attachement symbolique à l'intérêt économique dans certains cas. La relation individuelle peut être travaillée par les plateformes de services pour qu'un usage s'installe et devienne productif à leur avantage. La connaissance produite comporte un volet lié aux caractéristiques de l'usager lui-même, mais pas exclusivement. Il est, dans ses contributions spontanées et ses réactions, une force collaborative de

[98] La métaphore des "lignes" qu'il faut démêler est une reprise du vocabulaire de G. Deleuze (Deleuze, 1989, p.185). Il en va ainsi des lignes de visibilité, de forces ou de fuite.

production. Nous retrouvons dans ce mécanisme les caractéristiques de dispositif que nous avons soulignées tout au long de ce chapitre.

De manière évidente, ces dispositifs unitaires ont à leur tour vocation à être articuler dans des dispositifs complexes. De ce point de vue, le terme d'agencement, qui renvoie à une réalité essentiellement d'ordre spatial, ne porte pas en lui-même la dynamique et la plasticité que l'on peut attendre d'un dispositif. La continuité spatiale et la permanence du numérique nous amène à penser autrement qu'au travers d'objets, de lieux ou de temps dédiés. L'absence de stabilité du rapport des parties au tout, qui paraissait être un frein à l'emploi de la notion, s'avère au contraire un élément clef de sa définition. Il s'agit en effet d'intégrer la complexité structurelle du dispositif si l'on souhaite rendre cette notion productive.

Pour cela, nous préférons au terme agencement celui de *configuration* tel qu'il est envisagé par Norbert Elias[99] qui apparait beaucoup plus heuristique. Pour N. Elias, ce terme traduit bien l'idée d'une continuité qu'il estime nécessaire dans la manière de considérer les individus au sein d'un collectif et le collectif en tant que groupe. Introduire ce terme lui permet de se démarquer de la notion de structure trop connotée. Il adopte le terme de configuration plus dynamique et moins rigide que celui de structure : « *Il est plus commode de parler de configurations d'êtres humains, par exemple de la configuration mouvante que forment deux équipes de joueurs sur un terrain de football.* » (Elias, Dunning, 1994, pp.60-61). L'analogie sportive qu'il développe, souligne l'interdépendance entre les différents protagonistes du jeu. Elle met en évidence le fait qu'au cours de la partie, la configuration de jeu est conforme aux règles qui le régissent de même que chacun des comportements individuels. Chacune des configurations singulières et instantanées que l'on peut saisir est cohérente et traduit un équilibre dynamique des tensions internes à l'œuvre au sein (et entre) les différents regroupements d'acteurs pertinents.

L'équilibre dynamique de la configuration ne peut se laisser saisir que dans l'instantané. Caractériser un dispositif conduit alors à le décrire de manière récursive, comme un ensemble d'entités hétérogènes (parmi lesquels d'autres dispositifs) développant des configurations productrices de sens et agissantes.

La récursivité dans la définition introduit une structuration de configurations suivant différents niveaux de profondeurs, ce qui permet de mettre en perspective les contributions respectives et d'unifier des parties non développées. Ainsi, l'utilisation par une personne d'un dispositif info-communicationnel constitue en elle-même un dispositif. Caractériser un dispositif complexe ne peut raisonnablement pas s'envisager sans les coupes réductrices que définissent les différents niveaux. On peut alors envisager Twitter comme un tout ou le décomposer dans les différentes entités hébergeant des contenus (instagram, youtube, etc.) qui en assurent la consistance et la pertinence en tant que dispositif.

Dans le contexte empirique de nos études, identifier des dispositifs et des configurations comporte une part d'indétermination qui se lève progressivement en suivant les productions

[99] Nous renvoyons à la lecture de l'article d'André Ducret pour une analyse sociologique plus poussée de la notion chez Elias (Ducret, 2011).

informationnelles des entités principales (identifiées aisément). C'est par transitivité que nous arrivons à circonscrire le dispositif qui fera alors l'objet de l'observation.

Les relations que nous mettons ainsi en évidence ne couvrent donc pas l'étendue des relations existantes entre les entités mais seulement celles visibles (ou de surface) qui contribuent à l'effectuation de services étudiés. Cette restriction fixe les limites de l'observation numérique qui doit être complétée par d'autres méthodes d'investigation si l'on souhaite comprendre les logiques internes des dispositifs. Néanmoins, la production et la circulation d'information entre les entités est selon nous une caractérisation suffisante justifiant l'analyse des usages au travers des traces numériques que ces flux engendrent dans l'espace numérique.

Enfin, de manière globale, la définition de dispositif nous parait adaptée avec une approche analytique systémique que nous privilégions dans nos travaux. En particulier, les différentes observations réalisées notamment dans le contexte politique nous conduisent à raisonner des conditions d'équilibres atteintes dans les termes de l'*homéostasie* des systèmes complexes fermés. Bien que cela fasse écho à certaines de nos intuitions, nous hésitons encore à faire référence au concept d'*autopoïèse* dans nos analyses des dispositifs info-communicationnels tels que Twitter. Cela nous parait être une piste d'exploration à envisager dans certaines circonstances de structuration collective et affinitaire. Une telle exploration nécessiterait cependant que nous soyons en mesure de caractériser des représentations complexes et d'en étudier l'évolution. Les conditions de caractérisation, de représentation et d'analyse ne sont pas encore réunies pour de telles investigations qui fixent cependant un horizon à nos travaux.

CHAPITRE 2
Méthodologies de l'usage

Pour rappel, nous nous restreignons aux configurations de dispositifs dans lesquelles des entités numériques qui les constituent sont en capacité d'assurer au moins un service ou une médiation de service, de nature informationnelle ou communicationnelle. Dans ce chapitre, nous conservons les définitions de dispositif ou de médiation associées aux modèles que nous présentons.

Pour tenir le présupposé méthodologique que nous nous sommes fixé, il ne nous est pas apparu souhaitable de refermer le présent chapitre sur la question de l'*usage* du point de vue des SIC ou des SHS. En effet, la réflexion sur les pratiques sociales des TIC que l'on désigne dans les SIC comme *la sociologie des usages*, n'épuise pas l'intérêt pour les *traces d'usage*. La perspective que nous adoptons vise la possibilité, au-delà de l'étude des interactions et des médiations numériques situées, d'une analyse de comportements et de phénomènes sociaux dépassant le registre instrumental de l'activité. Cette ouverture permet d'associer des réflexions d'une autre nature sur la conduite instrumentée et sur la finalité de l'action, favorisant ainsi les collaborations inter disciplinaires et l'inscription de nos travaux dans un champ plus large que celui d'une sociologie des usages.

Si la sociologie des usages n'est pas ici une fin en soi, ses approches méthodologiques peuvent nous guider dans l'exploration que nous entreprenons. Par ailleurs, le cadre computationnel que nous adoptons, c'est-à-dire le fait de mettre en regard des modèles analytiques et des traitements informatiques, contribue à la définition d'un processus interprétatif par étapes. Cette perspective permet de structurer des niveaux analytiques dans une progression compatible avec les enjeux de l'usage.

L'objectif de ce chapitre est d'identifier les perspectives méthodologiques principales adoptées dans les approches de l'agir instrumenté et de l'usage des TIC. Pour cela, nous envisageons les questions du devenir instrument de l'outil (instrumentalisation) et de son incorporation dans les logiques d'activité (instrumentation). Partant de ce point de vue (§1), proche de préoccupations psychologiques sur le *sujet*[100] agissant, nous abordons les fondements et les méthodologies de l'analyse de l'usage dans un sens étendu (§2). Si la première partie nous situe dans la dynamique d'une confrontation de l'individu à la technique, la seconde rompt avec le sujet épistémique pour saisir l'individu dans des pratiques situées. Ce changement de perspective permet d'envisager les contraintes structurelles et configurationnelles constitutives du dispositif dans lequel l'usage

[100] Le *sujet* désigne ici un utilisateur faisant l'objet d'une observation. Le terme est utilisé dans un sens psychologique. Il peut être associé à un utilisateur occasionnel ou usager régulier.

s'élabore. Nous pourrons alors dans la suite de ce mémoire (partie II), situer nos travaux et envisager dans quelle mesure et de quelle manière, les techniques et les méthodes de la fouille des données d'usage s'inscrivent dans - ou peuvent s'articuler avec – des modélisations établissant un pont avec l'activité et les finalités qui la gouvernent. Nous serons alors à la croisée de deux perspectives orthogonales. L'une va dans le sens d'une compréhension comportementaliste et individuelle, l'autre dans celui d'une lecture phénoménologique engageant l'individu dans un rôle socialement situé.

1. Agir instrumenté

Devant l'existence de multiples désignations possibles de l'objet technique ou artefact, nous le désignerons suivant le statut d'*outil* lorsqu'il est mobilisable dans une activité et celui d'*instrument* lorsqu'il est effectivement agi. Cette distinction s'inscrit dans la continuité des travaux d'André Leroi-Gourhan pour qui l'outil n'existe que dans le cycle opératoire (Leroi-Gourhan, 1965).

Évoquer la question de l'agir instrumenté apparaît comme un exercice délicat. Il s'agit en effet de ne pas revenir en arrière en adoptant un point de vue ou une problématique d'IHM. Ici l'horizon n'est pas celui de l'interface et l'ambition n'est plus de formaliser les mécanismes de l'interaction dialogique Homme-Machine afin d'améliorer les caractéristiques des systèmes interactifs. Les problématiques auxquelles nous nous confrontons vont désormais au-delà de l'outil et de son *utilisabilité*. Elles concernent ce qui est accompli avec ces outils suivant une définition de l'*utilité* au sens de *utility*[101] défini par J. Nielsen (Nielsen, 1994) et qui n'est pas celle envisagée par le concepteur ou *marchandisée* par son promoteur, mais celle révélée par les usages. En ce sens, nous restons (du moins dans un premier temps) sur des niveaux d'analyse et des domaines qui voisinent l'étude des interactions homme-machine.

1.1. L'utilité instrumentale

L'utilité n'est pas une propriété inhérente de l'outil (Blandin, 2002). L'utilité est, dans le cas présent, une évaluation de l'adéquation entre l'outil et la conduite de l'activité. Elle ne s'établit que de manière circonstancielle et dans la satisfaction de l'utilisateur.

Ce détour par l'utilité peut paraitre réducteur. En effet, raisonner sur des fins est souvent associé à une approche déterministe de l'interaction. Mais, en définissant l'adéquation en référence à l'outil de manière générale et non à sa finalité programmée, nous ne lions pas la finalité de l'utilisation à une prédestination quelconque de l'outil. Ce relâchement de contrainte sur l'usage normatif de l'instrument ouvre le champ de l'action instrumentée : on peut faire plus et autrement que ce qui est prévu.

Suivant ces hypothèses, la décision d'instrumenter l'action et la manière d'opérer sont plutôt considérées comme relevant de choix tactiques et circonstanciés, liés à l'initiative de l'utilisateur.

[101] Chez J. Nielsen *Utility* a un sens plus étroit que *Usefulness* (qui l'inclut avec l'*Usability*).

Le présupposé d'autonomie de l'utilisateur accompagne la référence à l'utilité. Cette hypothèse forte vient du fait que nos études ont principalement porté sur des contextes d'utilisation spontanée et non réglée (des contextes majoritairement associés au cadre privé, bien distinct de l'usage situé dans des contextes professionnels). L'autonomie que nous envisageons n'est toutefois pas celle d'un individu coupé des réalités sociales du monde : le jugement d'utilité n'est pas exempt d'influences externes et les contraintes sociales y sont opérantes.

La référence au discernement de l'utilisateur nous permet d'assumer également un présupposé de rationalité des comportements de l'utilisateur, y compris dans sa décision d'instrumenter son action (cf. §1.3). Ce portrait idéalisé d'un individu autonome et acteur, nous amène à formuler l'hypothèse d'une intentionnalité présidant aux décisions d'instrumentation et d'action.

Dans la plupart des travaux que nous avons conduits, la satisfaction de l'utilisateur est un présupposé que nous posons du fait de l'utilisation des outils à disposition, fait que nous entérinons par l'observation de traces d'usage. Cette hypothèse vient le plus souvent dans une analyse *ex post*. C'est alors l'ensemble des traces d'une activité terminée qui est pris comme indicateur d'utilité.

Le recours à la notion d'utilité nous permet d'assurer la pertinence *a priori* de l'approche via les traces d'usage de l'activité. Cette remarque appelle une précision. L'instrument est en effet un médiateur de l'action. En cela, l'utilité peut renvoyer à la médiation ou au résultat obtenu sans que l'on puisse distinguer clairement laquelle de ces facettes est déterminante. Bernard Blandin souligne que cette distinction constitue une difficulté, notamment en sociologie des usages, à penser les usages : «… *les usages des médias comme les usages des objets techniques sont considérés comme des activités ou des actes ayant pour finalité l'emploi du média ou de l'objet technique, alors qu'ils ne sont en réalité que la forme prise par une action dont la finalité est autre.* » (Blandin, 2002, p43).

Cependant, cette ambivalence de notion d'utilité exprime une unité globale qu'il nous parait nécessaire de conserver. Une manière de prendre en compte différents aspects liés à l'utilisation d'un artefact est d'employer la distinction introduite par Pierre Rabardel entre ce qui relève des logiques de l'*instrumentalisation* et de l'*instrumentation* de l'activité (Rabardel, 1995).

1.2. L'instrumentalisation

L'instrumentalisation est selon Pierre Rabardel un processus dirigé vers l'artefact. Ce processus mental est celui d'une élaboration progressive par le sujet d'une représentation enrichie des propriétés de l'outil au fil de son usage. Pour P. Rabardel, ce processus prend appui sur des propriétés intrinsèques de l'artefact auxquels l'utilisateur attribue en situation un statut en fonction de l'action engagée (*Ibid.*, p114). Ces affectations (sans transformation de l'objet) peuvent perdurer au-delà d'une situation particulière et devenir des propriétés extrinsèques permanentes de l'objet[102].

Le processus d'instrumentalisation est un processus dynamique, décrit comme très général et permanent. Ainsi, dans la terminologie que suggère l'auteur, l'instrument est un état momentané et situé, qu'acquiert l'outil dans les mains de l'utilisateur, c'est-à-dire en fonction de son expérience.

[102] C'est de cette manière qu'il illustre le détournement de la clé à molette devenant à l'emploi, un marteau. Les propriétés intrinsèques de masse, non déformation, etc. devenant une propriété fonctionnelle contondante extrinsèque.

Du point de vue de la psychologie et de la cognition, ce changement de statut de l'outil est associé à représentations mentales et des opérations qui conditionnent l'action individuelle. Le concept de *schème* est de ce point de vue incontournable. Selon P. Rabardel (*Ibid.*, p76) les schèmes organisent des connaissances de différentes natures ; un schème est « *un cadre assimilateur qui attribue des significations et qui exerce une fonction se réalisant essentiellement dans la planification.* » (*Ibid.*, p84). En particulier, les schèmes familiers structurent les acquis et contribuent à l'interprétation des situations nouvelles (fonction heuristique[103]). Suivant l'hypothèse de P. Rabardel, « *c'est l'association de schèmes familiers (schèmes d'utilisation) aux artefacts qui, en attribuant des significations aux artefacts, aux objets et à l'environnement est constitutive des instruments*[104] » (*Ibid.*, p85).

L'approche de Pierre Rabardel incite à ne pas considérer des Technologies info-communicationnelles comme réduites à des fonctions mais comme des outils ou des *machines à communiquer* pour reprendre les suggestions de Pierre Schaeffer et de Jacques Perriault[105].

1.3. L'instrumentation

L'instrumentation est un second processus instrumental qui, comme le précédent, s'inscrit dans une durée. Ce processus est mobilisé dans une situation de résolution de problème correspondant à un cas d'activité durant lequel l'outil (disponible) est susceptible de contribuer à sa réalisation. De la même manière que pour le processus d'instrumentalisation, ce processus dépend de la situation, et pour l'utilisateur, des schèmes familiers associés à l'activité ainsi que des schèmes issus de l'instrumentalisation.

Deux types de schèmes sont mobilisés à ce stade et constituent la classe des schèmes sociaux d'utilisation. (*Ibid.*, p91) : les *schèmes d'usage* qui s'instancient dans les cas assimilés d'utilisation de l'instrument ; les *schèmes d'actions instrumentés* qui « *sont constitutifs de ce que Vygotsky appelait les "actes instrumentaux", pour lesquels il y a recomposition de l'activité dirigée vers le but principal du sujet du fait de l'insertion de l'instrument.* ». Ces schèmes ont une réalité personnelle mais aussi sociale. Pour P. Rabardel, la socialisation des schèmes vient de ce que l'utilisateur n'est pas isolé dans ses pratiques et qu'il incorpore de ce fait des éléments de schèmes qui lui sont externes. Les schèmes d'utilisation remplissent trois types de fonctions contextualisées : épistémique, pragmatique et heuristique.

La répétition des situations, amenant des régularités ou des écarts faibles (assimilables aux situations normales) dans l'instanciation des schèmes sociaux d'utilisation conduit à leur généralisation et à leur renforcement. De façon symétrique, les écarts importants (non assimilables) conduisent à leur différenciation.

---

[103] C'est cet aspect heuristique qui fait que le sujet va plus vite dans la prise de décision, le choix d'un outil, d'une de ses fonctionnalités, quitte à l'utiliser de manière détournée. Il choisit la solution qui lui semble la solution la plus économique selon ce qu'il sait, et la plus efficace (théorie de la rationalité limitée de HA Simon).

[104] Cette définition est cohérente avec la définition que nous avons donnée de l'instrument dès lors qu'il est envisagé en potentialité d'action.

[105] Voire à ce sujet la remarque de Jacques Perriault sur la désignation des outils info-communicationnels (Perriault, 2015).

Comme le fait remarquer l'auteur, les outils contemporains intègrent des pratiques collectives et collaboratives. Il serait nécessaire de disposer d'un troisième schéma traduisant l'intégration de la dimension collective (*Ibid.*, p92) dans la définition des schèmes sociaux d'usage. Dans cette perspective, P. Rabardel donne de l'instrument une définition qui dépasse le cadre artefactuel en lui attribuant une nature double associant à la réalité objective et technologique, celle subjective qu'expriment les schèmes sociaux d'utilisation et leurs déclinaisons et adaptations situées. Cette double nature de l'instrument permet d'envisager différents mécanismes d'association fondés sur une continuité de pratiques et d'expériences qui est relative, suivant le cas, à la nature artefactuelle ou schématique.

Pour autant, la distinction précédente, instrumentalisation vs instrumentation n'est pas aussi tranchée. Il n'est pas évident d'articuler suivant ce couple des activités centrées sur l'artefact lui-même visant son paramétrage ou sa personnalisation et ces activités ressortent autant de l'instrumentalisation que de l'instrumentation.

1.4. Présupposés

Dans la définition de l'agir instrumenté, nous ne coupons pas à l'idéalisation de l'individu. Cette idéalisation se porte sur l'engagement individuel dans l'action et dans les éléments de détermination qui conduisent à son instrumentation. Cette représentation de l'individu repose sur un ensemble de présupposés sur des états mentaux associés aux comportements que l'on ne peut pas ignorer.

1.4.1. Rationalité

La rationalité de l'individu agissant est un présupposé qui a conduit à croire que l'utilisateur d'un dispositif info-communicationnel allait en faire ce que l'on attendait de lui. La rationalité est avant tout un jugement de conformité, établi dans l'ajustement des fins et des moyens. Dans le cas d'un outil, cet ajustement est posé *a priori* durant sa conception. L'usage conforme est celui envisagé par le concepteur. L'exemple du Minitel a clairement démontré que l'usage effectif était difficile à anticiper. Pour autant, l'usager bricoleur et inventeur de ses usages, n'est pas irrationnel. Les *catachrèses* instrumentales, c'est-à-dire le détournement d'usage est un phénomène qui doit être pris en compte. Selon P. Rabardel, cette propension à inventer l'usage est une forme de production de l'instrument et « *plus généralement des moyens de ses actions* » (*Ibid.*, p100).

Cette extension du domaine de rationalité aux usages, rejoint l'approche encouragée par Michel De Certeau. Elle est intégrée dans la démarche d'ethnotechnologie proposée par Jacques Perriault (cf. §2.1). Cependant, la portée de cette définition est limitée aux effets immédiats de l'instrument qui ne recouvrent pas les conséquences en situation et principalement sociales de l'action. La rationalité de l'agir instrumenté s'étend au-delà du périmètre de l'outil. Elle ne doit pas être confondue *a priori* avec la *rationalité instrumentale* qui désigne l'utilisation de la raison comme instrument. Selon Max Weber, la rationalité instrumentale, gouverne les actions déterminées : « *par des expectations du comportement des objets du monde extérieur ou de celui d'autres hommes, en exploitant ces expectations comme conditions ou comme moyens pour parvenir rationnellement aux fins propres, mûrement réfléchies qu'on veut atteindre* » (Weber, 1971, p55). Cette définition de la rationalité caractérise la mise en œuvre des

moyens en raison des fins. Nous ferons l'hypothèse que cette définition, plus générale car non restreinte au cas de l'action instrumentée, englobe la précédente et nous l'adopterons comme telle. La rationalité instrumentale n'est pas seule déterminant des comportements des usagers. Ceux-ci intègrent dans leur décision d'agir des normes sociales et des éléments d'une rationalité axiologique (éthique, etc.) (*Ibid.*).

1.4.2. Intentionnalité

Nous avons évoqué le concept d'intentionnalité dans le chapitre précédent afin de rendre compte d'un processus d'élaboration le plus souvent faiblement coordonné entre de multiples acteurs. Nous sommes désormais du côté de l'utilisateur, supposé rationnel. La rationalité instrumentale de Weber, parce qu'elle prend en compte la situation dans son élaboration, nous est apparue une modélisation nécessaire. Chez M. Weber, toute décision d'action, relevant d'une logique de moyen, est par nature intentionnelle. En revanche, elle n'est rationnelle que si les moyens sont en adéquation avec les buts visés. L'intentionnalité traduit une capacité du sujet à projeter son action et de manière plus générale à planifier son agir.

Dans les termes de l'interaction homme-machine, (IHM), rationalité et intentionnalité sont liés dans les modèles de description des tâches instrumentées calquées sur les préconisations d'usage des applicatifs. Les interfaces de dialogue s'appuient sur des modélisations de ce type caractéristiques de situations normales d'utilisation. Les modèles de croyance (*belief*) et d'intention (*intend*) inspirés des travaux de John Rogers Searle[106] (Searle, 1976) ont contribué à la mise en œuvre de moteurs de planification opérationnalisant la modélisation en tâches de l'activité dans les systèmes dialogiques. Le travail formel qui préside à ces modèles accrédite des mécanismes d'intelligence artificielle (IA) dont le statut cognitif n'est pas avéré. Ces modèles de l'IA ont largement influencé les modélisations en IHM, même au-delà du cadre d'activités prescrites.

2. Fondements et méthodologies d'analyse de l'usage

L'intérêt pour l'usage et l'usager des technologies naît avec le renouvellement théorique dans l'approche des usages médiatiques, dont émerge (vers 1940) la théorie dite des usages et gratifications (*uses and gratifications*).

Définir le concept d'usage en SHS c'est courir, dans notre champ disciplinaire, plusieurs dangers. Celui d'engager la nième tentative de synthèse historique, au risque de se perdre dans les méandres des apports disciplinaires très nombreux et de la synthèse improbable de travaux pléthoriques et faiblement connectés les uns aux autres. C'est encore courir le risque, non moins grand, de ne pas éviter les lieux communs tant dans le discours que dans les références scientifiques susceptibles de le fonder. Fabien Granjon et Julie Denouël dressent dans le chapitre introductif de l'ouvrage collectif qu'ils ont dirigé un tableau très clair de cette situation (Denouël, Granjon, 2011).

[106] Ces travaux se rapportant aux actes illocutoires, ont connu un très grand intérêt dans les communautés des IHM. Malheureusement, ces réflexions théoriques, souffraient d'un défaut d'ancrage linguistique que les travaux au sein du CRISS ont largement démontré.

Le numéro 6 de *Revue Française des Sciences de l'Information et de la Communication* (RFSIC) publié en 2015 dresse un point étape pour les travaux consacrés aux *Usages et Usagers de l'information à l'ère numérique* (Badillo, Pélissier, 2015). Nous saisissons cette actualité éditoriale comme cadre de lecture général de la problématique d'usage en SIC, nous accordant ainsi la liberté de ne pas engager un tour d'horizon exhaustif de la question au bénéfice d'une lecture critique de contributions en lien avec le présent mémoire.

Il est probable qu'en portant notre attention sur les « *usages de l'usage* » (Jeanneret, 2007) ou plus précisément les « *usages de la notion d'usage* » (Lacroix & Al., 1992) ce parcours rétrospectif n'aurait pas manqué de souligner, comme ces auteurs l'ont fait, que le terme d'usage constituait un « *mythe, c'est-à-dire un signe qui plane dans le discours, fortement coupé de son histoire et de sa genèse »* (Jeanneret, 2007).

En même temps, le terme d'usage n'est pas dénué d'intérêt. Davantage que le terme de pratique avec lequel il se confond généralement, il est porteur d'une continuité faisant écho, par son lien avec la tradition, à une certaine stabilité normative et culturelle. En cela, l'usage des TIC établit des passerelles avec l'anthropologie des techniques, ménageant ainsi plusieurs niveaux de lecture nous permettant d'embrasser, sous un même terme, des enjeux d'analyse et d'échelle variés.

Par ailleurs, enfermer l'*usage* dans une définition trop statique constituerait une erreur. Qu'il s'agisse de considérations sociales ou individuelles, l'usage n'est pas fixe, il est sujet d'adaptations ; « *…l'usage est toujours pluriel : il est pérennisé et inscrit dans des formes, mais aussi actualisé et porté par des pratiques vivantes.* » (*Ibid.*).

2.1. L'ethnotechnologie

L'ethnotechnologie est un néologisme forgé en mars 1976 pour désigner un groupe de travail qui s'est constitué à l'initiative du service technologie au ministère de l'industrie (Gaudin, 1981). L'objectif de ce groupe de recherche-action, composé de chercheurs d'horizons divers[107] et de fonctionnaires ministériels, était de réfléchir à la place de la technique dans la société à une époque où « *la technique est désormais perçue comme un dangereux mécanisme d'asservissement des hommes »* (*Ibid.*, p119).

Dans le document historique que constitue le texte de Thierry Gaudin, ce dernier utilise un schéma fort simple de ce qui cristallisa les débats dès les premières réunions de ce groupe de travail.

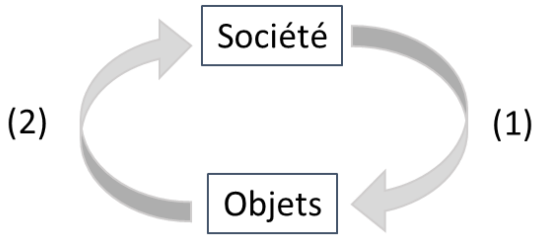

Fig14. Schématisation de l'effet des techniques

La branche (1) est associée à la production des objets, alors que la branche (2) manifeste la transformation de la société et des mœurs (*Ibid.*, p120). Si la branche (1) traduit une lecture

[107] On retrouve parmi eux des chercheurs dont Patrice Flichy et Jacques Perriault qui auront une influence dans le développement de la réflexion sur les TIC dans le champ des SIC en France.

industrielle et productiviste (légitime du point de vue de l'auteur), la seconde introduit un rapport qui n'est pas celui de la consommation mais celui d'un effet rétroactif (en conséquence) de la production sur la société.

Cette hypothèse modifie radicalement la portée du schéma ci-dessus. Une lecture classique aurait vu dans ce schéma le cycle d'entretien du marché (renouvellement de l'offre justifié par le besoin) sans effet sur la société. En cela, ce schéma devient dynamique et symbolise l'émergence d'une pensée socio-technique fondée sur l'idée d'une circularité amenant une mise en tension de la technique et du social.

Cependant, plusieurs lectures de ce cycle se superposent. Dans le cadre présent, deux sont mises en avant. La première privilégiant (1) est celle qui voit dans l'artefact la traduction et le véhicule d'une connaissance technique inscrite dans la société. Le cycle se boucle alors sur l'hypothèse d'une réalisation du progrès, transformation positive de la société par la technique. La seconde privilégiant (2), renverse la perspective. La transformation est celle des pratiques sociales issue de l'appropriation de l'artefact. La boucle se referme alors suivant l'hypothèse d'une (re)définition de l'objet dans l'usage.

Jacques Perriault a joué un rôle essentiel dans l'émergence de cette réflexion. Son œuvre s'inscrit naturellement dans la filiation de l'ethnotechnologie qu'il assume avec constance (Perriault, 2015). L'ouvrage majeur qu'il a publié en 1989 : *la logique de l'usage – essai sur les machines à communiquer* (Perriault, 1989), a fortement influencé l'approche des TIC au sein de notre discipline. Sans doute a-t-il contribué à légitimer le concept d'usage dans les SIC et par la même, autoriser une réflexion plus large sur les machines à communiquer. Cet auteur définit sa démarche de sociocognitive et d'anthropologique et la distingue de la *sociologie des usage*s. C'est dans cette voie distincte, qu'il qualifie d'orthogonale vis-à-vis de la sociologie des usages, qu'il faut considérer sa réflexion sur la *logique de l'usage* (Perriault, 2015).

2.1.1. La logique de l'usage

La définition de l'usage que propose J. Perriault (Perriault, 1989) doit être replacée dans un contexte historique où les discours sur les *nouvelles technologies* opposent des imaginaires du progrès (Scardigli, 1989) et des idéologies antagonistes. L'approche des technologies par l'usage constitue donc une rupture d'ordre méthodologique. Cette rupture découle du focus mis sur la relation d'utilisation et la volonté d'objectivation qui l'accompagne.

Une telle approche convoque, pour une durée finie, le sujet (individu), acteur central, et l'objet technologique (ici une machine à communiquer). La relation d'utilisation est par nature singulière et orientée. Elle relie dans un contexte donné, un individu animé d'un projet à une machine à communiquer auquel l'utilisateur attribue une fonction. Pour J. Perriault, le triplet Projet – Instrument – Fonction permet de décrire l'ensemble des ajustements pratiques opérés par l'individu (Perriault, 2015).

Pour autant, l'utilisation n'est pas l'usage. Celui-ci s'élabore dans la durée, la répétition et par ajustements négociés au fil des utilisations. L'usage personnel s'installe progressivement (ou non) comme un point d'équilibre dans le champ conflictuel existant entre «…*l'homme porteur de son projet*

et l'appareil, porteur de sa destinée première. » (*Ibid.*, p220). Les pratiques résultant de cette acculturation instrumentale ne couvrent qu'en partie (le plus souvent) le seul cadre d'action prévu. De ce point de vue, l'usage est par nature sous optimum. En revanche, l'appropriation non contrainte des potentialités instrumentales permet l'expression d'une créativité dans l'utilisation et étendre l'usage (voire le détourner) en dehors des cadres prévus. Cette capacité au détournement (ou catachrèse), fait écho aux analyses de Michel De Certeau (Certeau, 1990). Ces aspects non conventionnels de l'usage ont été depuis largement relayés. Ils ont ainsi humanisé les discours sur les objets technologiques et la technique qui, jusque-là, étaient considérés comme les moyens de la surveillance et de l'aliénation individuelle. Il ne faut cependant pas perdre de vue, que la négociation de l'usage n'est pas sans effets normatifs sur les pratiques de l'usager qui incorpore pour partie au moins les règles d'utilisation de la machine.

Dès lors qu'il est installé, l'usage conduit à des régularités comportementales individuelles qui s'agrègent à leur tour en régularité au sein de collectifs voire de la société dans son ensemble.

J. Perriault souligne également que la fonction instrumentale n'est pas restrictive. L'objet peut être investi symboliquement et l'usage acquérir une fonction rituelle. Mais là encore, les possibles sont réduits et des régularités individuelles et collectives se font jour.

Le temps long de l'usage correspond à la double réalité de l'incorporation individuelle des techniques et de la diffusion dans la société de cette appropriation. C'est dans cette convergence, qui ne peut s'observer que de haut, que l'auteur formule l'hypothèse de l'existence d'une logique sous-jacente à l'usage. L'analyse historique et anthropologique que suggère J. Perriault, témoigne de durées variables qui ne rendent pas seulement compte de déterminants cognitifs ou sociaux : le niveau de connaissance technique n'est pas seul en cause et d'autres facteurs individuels et sociétaux interviennent, tels que les mythes par exemple.

2.1.2. Vers une approche empirique de l'usage

En introduisant l'utilisation comme élément atomique de l'usage, l'ensemble des indicateurs globaux liés à la production ou à la consommation (taux d'équipements, etc.), issus du marché sont invalidés : posséder n'est pas user-utiliser. De la même manière, le réalisme associé à l'utilisation effective disqualifie tous recours aux référentiels *a priori* tels que les modes d'emploi et autres prescriptions d'usage comme significatifs de pratiques.

Dès lors, coupé de sources externes d'information, l'usage ne peut s'aborder que de manière empirique et expérimentale. La subjectivité des discours sur les pratiques personnelles, y compris dans le déroulé de l'action, a bien été mise en évidence. En conséquence, l'observation est une nécessité que l'œuvre de Jacques Perriault a clairement explicitée.

2.2. La Théorie de la structuration adaptative (AST)

La théorie de la structuration adaptative (*Adaptative Structuration Theory*) a été proposée au début des années 1990 dans une succession de travaux initiés par Wanda Orlikowski puis enrichis par Gerardine De Sanctis et Marshall Poole, qui établiront la théorie (Orlikowski, 1992), (Sanctis, Poole, 1994).

La réflexion de ces auteurs se situe dans le contexte de l'informatisation des organisations qu'ils abordent du point de vue stratégique et managérial. Sur le plan théorique, leurs travaux s'appuient sur le paradigme structurel, ce qu'ils justifient en référence à P. Bourdieu[108] dans *esquisse d'une théorie pratique* (Bourdieu, 1972) et de manière plus explicite en référence à la *théorie de la structuration* d'Anthony Giddens (Giddens, 1979).

L'objectif pour ces chercheurs est de comprendre les mécanismes d'appropriation des technologies mais aussi de leurs effets sur l'organisation elle-même. Le cadre organisationnel présente des caractéristiques et un dimensionnement humain qu'il est intéressant de rapprocher (toutes choses égales par ailleurs) des dispositifs info-communicationnels et communautaires que nous étudions. De plus, l'étude des systèmes d'information implique une démarche de formalisation des processus info-communicationnels qui est très similaire à la nôtre.

2.2.1. Modèle structurationniste des technologies[109] d'Orlikowski

Wanda Orlikowski porte son attention de façon plus globale sur la manière dont les Sciences de gestion[110] envisagent l'incidence des technologies de l'information communication dans les organisations. Partant d'une critique des approches existantes à l'époque (début des années 1990), W. Orlikowski se propose d'aller au-delà du diagnostic des limites imputables aux perspectives (*scope*) et aux rôles attribués aux technologies dans les transformations organisationnelles. Il lui paraît indispensable d'avoir une perspective permettant d'atteindre le niveau micro des interactions individuelles et cela afin de prendre en considération les transformations sur toute l'étendue de l'organisation. Ceci suppose de dépasser le concept globalisant de technologie qui, dans l'emploi générique qui en est fait, neutralise les notions de tâche, de technique, de connaissance et d'outils, et empêche de confronter ces notions aux agents humains (*Ibid.*, p399).

C'est sur cette base qu'W. Orlikowski propose un modèle *structurationniste* des technologies (*structurational model of technology*). Comme elle le fait remarquer, les travaux d'Anthony Giddens ne traitent pas explicitement la question de la technologie dans le paradigme structurationniste. Elle constate par ailleurs que, si la théorie structurationniste a déjà été envisagée, notamment par G. De Sanctis et M. Poole, dans des travaux abordant le rôle des TIC sur les organisations humaines, aucun de ces travaux n'a essayé d'étendre la proposition théorique de A. Giddens aux technologies.

Wanda Orlikowski introduit deux prémisses qui fondent sa modélisation. Elle désigne la première comme celle de la *dualité des technologies (the duality of technology)*. Cette prémisse traduit ce qu'elle conçoit comme une récursivité dans le fait que : « *Technology is created and changes by human action, yet*

---

[108] Dans la traduction anglaise publiée en 1978 sous le titre : *outline for a theory of practice* Cambridge University Press. La référence n'apparaît que chez G. De Sanctis et M.Poole.

[109] Problème de traduction de Technology en français. Nous choisissons d'utiliser le "pluriel" lorsqu'il s'agit d'évoquer le terme de technologie en tant que classe.

[110] Ou Sciences du management ?

it is also used by humans to accomplish some action.[111] » (*Ibid.*, p405). Cette lecture récursive de la technologie, empruntée de la théorie de A. Giddens, pose que les activités sociales des acteurs sont récursives, qu'elles se régénèrent en permanence dans l'usage qu'ils font des moyens (ici les technologies), moyens que ces activités ont-elles-mêmes contribuées à définir (Giddens, 1984).

La seconde prémisse est corollaire de la première. Elle la désigne sous l'intitulé l'*interprétativité flexible des technologies* (*the interpretively flexible*). Cette prémisse signifie que l'interaction entre technologies et organisations est de nature contextuelle, c'est-à-dire dépendante, des acteurs et des contextes socio-historiques impliqués dans le développement et l'usage des technologies. Là encore, la contextualité est une référence à la théorie d'Anthony Giddens.

Le modèle structurationniste des technologies se construit sur trois entités : les agents humains (les concepteurs, les usagers, les décideurs), les technologies (les artefacts contribuant à l'exécution de la tâche dans l'espace de travail), les propriétés institutionnelles. Cette dernière entité correspond, suivant la définition qu'en donne l'auteur, à un catalogue de descripteurs de l'organisation que nous livrons conforme à l'origine :

« *structural arrangements, business strategies, ideology, culture, control mechanisms, standard operating procedures, division of labor, expertise, communication patterns, as well as environmental pressures such as government regulation, competitive forces, vendor strategies, professional norms, state of knowledge about technology, and socio-economic conditions.* »[112] (*Ibid.*, p409).

Cette compréhension contextualisée de l'institution permet de s'approcher de la définition de *dispositif* telle que nous l'avons envisagée dans le premier chapitre. Le modèle structurationniste se présente alors ainsi (Orlikowski, 1992, p410) :

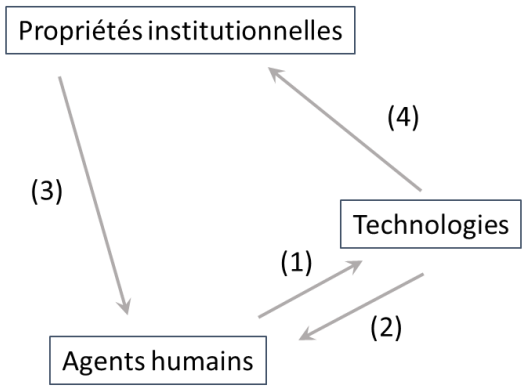

Fig15. Modèle structurationniste

[111] « La technologie est créée et modifiée par les actions humaines, mais elle est aussi utilisée par des humains pour accomplir des actions. »

[112] « Les arrangements structurels, les stratégies d'activité (business) l'idéologie, la culture, les mécanismes de contrôle, les procédures opératoires standards, la division du travail, l'expertise, les schémas de communication, autant que les pressions environnementales telles que la régulation gouvernementale, les forces en compétition, les stratégies de vente, les normes professionnelles, l'état des connaissances technologiques et les conditions socio-économiques. »

Flèche	Type d'influence	Nature de l'influence
(1)	Les technologies comme produit	Les technologies sont le résultat d'actions humaines de conception, développement, appropriation et modification
(2)	Les technologies comme medium	Les technologies facilitent et contraignent les actions en fournissant des schémas interprétatifs, des aménagements et des normes
(3)	Conditions institutionnelles de l'interaction	Les propriétés institutionnelles influencent les individus dans leurs interactions avec les technologies, par exemple dans les intentions (d'usage), les normes professionnelles, les états de l'art concernant les matériels et les connaissances, les standards de conception et l'affectation des ressources (compétences, financières, et temporelles).
(4)	Conséquences institutionnelles de l'interaction	L'interaction avec les technologies influence les propriétés institutionnelles d'une organisation, en renforçant ou transformant les structures de signification, de domination et de légitimation.

Fig16. Modèle structurationniste - description

L'originalité de cette schématisation réside dans la double influence des technologies, sur les agents humains suivant les cycles d'usage (1)(2) ou d'institutionnalisation (1)(4)(3). La transposition de ce modèle issue des organisations à la société fait resurgir la difficulté de distinguer les entités agissantes et de signifier leur interdépendance au sein d'une même configuration.

2.2.2. La formulation théorique de la structuration adaptative

Gerardine De Sanctis et Marshall Poole poussent plus avant la réflexion engagée par Wanda Orlinkowski dans l'adoption des théories d'Anthony Giddens (Sanctis, Poole, 1994, 1998, 2004). La prémisse de *dualité des technologies* est exprimée plus nettement comme une dualité de structures suivant laquelle il y a une interaction entre les types de structures inhérentes aux technologies (avancées) et les structures qui émergent dans l'interaction avec ces technologies.

Il convient de préciser que ces auteurs travaillent dans le cadre particulier des logiciels d'aide à la prise de décision collective (*Group Decision Support System* ou GDSS). Leur approche est aussi caractérisée par une référence aux technologies informationnelles (IT) avancées. La distinction avancée est historique et caractérise l'émergence des plateformes collaboratives dans le domaine du GDSS dans les années 1990. Depuis, cette distinction s'est généralisée et n'a plus besoin d'être soulignée. Les logiciels qu'ils évoquent ont très naturellement migré vers les technologies du Web et se conforment aux architectures client/serveur des plateformes de services que nous avons pris comme archétype de technologies info-communicationnelles (TIC).

Suivant leur formulation, les développeurs de logiciels applicatifs info-communicationnels intègrent, en les reproduisant, des éléments issus des structures sociales. Ces structures mimétiques

sont de ce fait mobilisées au cours des interactions avec ces applicatifs (ou services). Ces structures incorporées dans les technologies (contenus dans l'applicatif) s'entremêlent avec les structures sociales inhérentes aux groupes mobilisés contribuant, selon eux, à une mise en relation récursive des technologies et de l'action (Sanctis, Poole, 1994, p125).

Plus précisément, dans leur approche des structures sociales véhiculées par les technologies, les *caractéristiques structurelle*s (*structural features*) désignent le type de règles et de ressources spécifiques, les potentialités fonctionnelles proposées par le système. Il s'agit, selon nous, des déterminants techniques objectifs, établissant la spécificité et la distinction de technologies (*i.e.* service) parmi d'autres.

L'intentionnalité ou *esprit,* selon l'expression des auteurs, correspond aux buts généraux et aux attitudes que promeuvent les caractéristiques structurelles. Il est le principe de cohérence qui sous-tend les caractéristiques structurelles et leur implémentation. *L'esprit* est un concept par nature sous-déterminé. Selon ces auteurs, il est à saisir par analogie avec l'expression l'*esprit des lois*. Ainsi, on peut comprendre, que les caractéristiques structurelles sont élaborées dans un esprit global qui peut être contrarié dans l'utilisation du fait de l'indépendance fonctionnelle qui est un principe d'implémentation.

Les auteurs associent le terme de *structuration* au processus d'adaptation mis en œuvre par les groupes qui élaborent leur structure de travail à partir des caractéristiques structurelles des technologies mobilisées dans leur activité. M. Poole et G. De Sanctis relient cette définition au concept de *système* chez A. Giddens : « …*structuration can be defined as the process by which systems are produced and reproduced through members' use of rules and resources* »[113] (Poole, Sanctis, 1998, p4). Ils précisent bien que les deux structures n'ont pas de réalité indépendante, du moins en dehors des pratiques sociales qu'elles constituent.

Ce processus de structuration est constamment actif, c'est-à-dire susceptible d'aboutir à des structures différentes. La stabilité des structures dépend de différents facteurs et dynamiques externes et internes aux groupes.

Remarquons, que la finalité (dans ce cas, il s'agit d'aide à la décision) et les situations de références (l'organisation) sont des spécificités de leur travaux qui ne semblent pas être des éléments déterminants dans la formulation qu'ils proposent de la théorie. Les contraintes prescriptives et les logiques d'accompagnement des pratiques mises en œuvre dans les organisations agissent certes sur la structuration, mais elles ne sont pas immuables et leurs effets ne sont pas susceptibles de perdurer au-delà d'une période d'appropriation[114]. Le contexte des GDSS[115] est pour eux un avantage (mais pas une restriction) qui tient dans le fait que les structures sont très saillantes et donc aisément repérables et analysables (Sanctis, Poole, 1994, p143). Par nature, le GDSS mobilise le groupe dans son entièreté. La pratique individuelle est subordonnée à une pratique collective : la

---

[113] « La structuration peut être définie comme le processus par lequel les systèmes sont produits en reproduits au travers de l'usage par les membres des règles et des ressources. »

[114] C'est l'une des caractéristiques d'autonomie des acteurs soulignée par A. Giddens.

[115] *Group Decision Support System*.

décision d'instrumenter l'activité si elle n'est pas imposée, relève d'une logique collective même si les jeux de pouvoir et les rapports d'autorité n'en sont pas exclus.

Dans les situations que nous étudions, la décision d'usage est *a priori* considérée comme individuelle, sans pour autant négliger l'influence contingente du groupe et des circonstances événementielles auxquels l'utilisation raccorde l'usager. C'est donc davantage dans le déroulement de l'utilisation que la nature intrinsèquement collective (mais pour des groupes de taille réduite) et collaborative des mécanismes de structuration que l'on peut rapprocher nos préoccupations (cf. partie 2).

2.2.3. Les apports méthodologiques

Le cadre de réflexion de ces différents auteurs est lié aux systèmes d'information des organisations et à leur management. Les modélisations qu'ils proposent ont pour objectif de mieux comprendre les mécanismes d'appropriation des technologies au sein des organisations mais aussi une finalité managériale et pratique : pouvoir mieux les définir et en accompagner les pratiques. Dans cette perspective, la réflexion théorique se double d'une réflexion méthodologique dans la manière d'exploiter la modélisation (*i.e.* l'AST) pour analyser l'incidence des technologies et les transformations organisationnelles induites. Ces chercheurs estiment par ailleurs, que leur expérience méthodologique est transposable dans d'autres contextes technologiques et d'autres configurations sociales. L'intérêt pour nous est évident.

G. De Sanctis et M. Poole propose une stratégie d'analyse diachronique du processus de structuration en dix points dont les cinq premiers concernent l'appropriation des technologies (Fig17) ; les suivants se rapportent spécifiquement à l'applicatif (GDSS) et à ses conséquences sur l'activité (décisionnaire). Dans leur contexte, le groupe est une variable d'analyse permettant des comparaisons (synchroniques) entre différents groupes constitués. Les comparaisons peuvent également confronter des cas d'usage et des situations équivalentes mais non instrumentées.

étape	description
(1)	Décrire les caractéristiques structurelles et l'esprit des technologies mobilisées. (documentation technique, entretiens usagers, entretiens concepteurs, observations, etc.) doit être plus qu'une description fonctionnelle ou de l'interface. Elle doit permettre la mesure et la comparaison.
(2)	Identifier et décrire les autres sources de structures agissantes. Suivant les mêmes préconisations. Cette description peut porter sur les tâches à accomplir (richesse, complexité, etc.). il s'agit de pouvoir mettre en évidence le degré d'ajustement (*fit*) des technologies à leur environnement.
(3)	Décrire la composition du groupe en tant que système. (compétences individuelles, expérience collective, relations d'autorité, degré d'adhésion au projet instrumental.)
(4)	Formuler des hypothèses concernant l'appropriation des technologies
(5)	Évaluer l'étendue de l'appropriation, du degré de conformité des utilisations, l'usage instrumental, et l'attitude à l'égard de l'appropriation.

Fig17. Modèle De Sanctis et Poole d'analyse diachronique

Les deux auteurs inscrivent leurs travaux dans la tradition structuraliste. Les discours produits par les différents acteurs constituent la principale source de l'analyse. L'approche linguistique les conduit également à s'intéresser à la structure des échanges et aux actes de paroles (*speech acts*) produits au sein du groupe.

L'importance de la verbalisation peut apparaître problématique compte tenu de ce que nous avons rapporté des travaux de J. Perriault. Les circonstances diffèrent cependant. Dans le cadre présent, les énonciations se rapportent principalement à la mise en œuvre commune et coordonnée du système de décision. Les échanges traduisent les ajustements entre les membres impliqués dans ce processus. C'est à ce niveau que se situe l'intérêt pour une pragmatique linguistique : plus les discours sont nourris et complexes et plus on s'éloigne de l'usage ordinaire (stable et incorporé). Cette perspective méthodologique n'est pas sans lien avec celles de l'analyse des situations de résolution de problèmes étudiées en psychologie et ergonomie cognitive.

2.3. La sociologie des usages, l'heure d'un bilan ?

Des auteurs comme Josiane Jouët, Geneviève Vidal ou Serge Proulx, contribuent activement et régulièrement à questionner la notion d'usage et à l'inscrire dans la discipline (Jouët, 2000, 2011), (Vidal, 2012), (Proulx, 2001, 2005, 2015). Leurs lectures rétrospectives mettent en évidence la structuration d'un domaine de réflexions, notamment autour d'une sociologie de l'usage, et les ramifications progressives de recherches qui se spécialisent tout en s'ouvrant à l'interdisciplinarité. Dans un récent article, Serge Proulx revient sur l'évolution des problématiques d'usage dans le champ des SIC (Proulx, 2015). Il souligne en particulier, le glissement opéré dans le milieu des années 1990, amenant d'un topique centré sur les objets techniques (TIC) à un topique centré sur l'activité « *La technologie devient une dimension de l'écologie humaine et sociale parmi d'autres* » (*Ibid.*). Dans l'historique qu'il trace et qu'il arrête le en 2010, le second topique est étroitement lié au développement d'une sociologie des usages.

L'ère nouvelle qui s'ouvre depuis semble appeler selon lui un questionnement que porte le titre de l'article : *La sociologie des usages, et après ?*. Pour S. Proulx, les raisons de cette question sont à rechercher dans le bilan dressé en 2011 par Josiane Jouët (Jouët, 2011). Cette dernière souligne la part croissante des travaux se rapportant au Web et à ses services dans le champ de la sociologie des usages. Cette orientation, légitime au demeurant, amène J. Jouët comme S. Proulx à formuler des craintes et des mises en garde qui traversent actuellement l'ensemble des SHS. Les auteurs attirent l'attention sur l'attrait des traces d'usage et l'*extrême quantification* qu'elles induisent dans le rapport au monde. Cette expression traduit une double réalité : celle d'une *data(i)fication*[116] c'est-à-dire d'une mise en données, à la fois numérisation et catégorisation, de l'individu et du social ; celle d'une *massification des données* appelant une médiation computationnelle dans l'analyse.

[116] Le terme contemporain de *datafication* a été popularisé par Kenneth Cukier et Viktor Schoenberger dans un article de 2013. Nous y reviendrons plus longuement au chapitre 6.

Le regard rétrospectif porté par ces différents auteurs met en évidence un questionnement méthodologique ouvert. J. Jouët ou G. Vidal partagent le constat d'une multiplication d'études de faible envergure, repliées sur des cas et un prisme *micro*. Elles soulignent l'absence d'une réflexion de nature théorique apportant au domaine les moyens de sa régénération mais aussi l'accumulation de données privilégiée au détriment de l'analyse. À ces critiques, on peut opposer un temps de maturation qui rend l'accumulation d'expériences nécessaire. Il en a été ainsi du champ d'études de Twitter pour lequel les premières études menées effleurent les questionnements sociologiques et privilégient les comptages et les statistiques élémentaires.

S. Proulx soulève quant à lui, les difficultés inhérentes à la prise en charge des traces d'usage : « *les corpus s'organisant autour des seules traces des utilisateurs risquent de conduire à un empirisme méthodologique à outrance sans consistance théorique, l'épaisseur sociologique des usages se réduisant à n'être plus qu'une comptabilisation de clics »* (Proulx, 2015).

Au travers de ce questionnement d'ordre méthodologique se pose celui du devenir d'un champ d'étude qui n'est pas celui d'une discipline constituée. En introduction nous avons évoqué la nécessité imposée dans nos travaux de dépasser le cadre de la sociologie des usages tout en essayant de rester méthodologiquement cohérent avec ces problématiques. C'est la raison pour laquelle il nous parait intéressant de revenir sur l'analyse de S. Proulx sur ce champ d'études. Cette présentation s'appuie sur deux articles *Penser les usages des TIC aujourd'hui : enjeux, modèles, tendances* (Proulx, 2005) et *La sociologie des usages, et après ?* (Proulx, 2015).

À dix ans d'intervalle, l'auteur poursuit une réflexion sur les fondements méthodologiques de la sociologie de l'usage (second topique). Il dégage ainsi 5 axes qui constituent les perspectives sur lesquelles se rejoignent différentes disciplines. Ces axes s'ordonnent de manière croissante.

Il en résulte le tableau suivant :

(1)	2005	*L'interaction dialogique entre l'utilisateur et le dispositif technique*
	2015	*Suivre l'utilisateur dans son face-à-face avec l'objet technique : décrire l'interaction dialogique* utilisateur / dispositif technique (*Human-Computer Interaction* - HCI) ;
(2)	2005	*La coordination entre l'usager et le concepteur du dispositif*
	2015	*Suivre le cours d'actions de coordination entre le concepteur et l'usager* : cet angle postule une perméabilité entre les univers du concepteur et de l'usager ; le concepteur inscrit des « scripts » (Akrich, 1987) dans les objets techniques, inscriptions corrigées et ajustées en permanence en fonction des attentes et des pratiques déployées par l'usager ;
(3)	2005	*La situation de l'usage dans un contexte de pratiques (c'est à ce niveau que l'on pourrait parler de l'expérience de l'usager) ;*

	2015	*Décrire de manière fine et détaillée la situation d'usage : décrire de façon étoffée (thick description) les pratiques des agents et des collectifs dans l'environnement équipé (description compréhensive de l'expérience de l'usager individuel ou collectif) ;*
	2005	*L'inscription de dimensions politique et morale dans le design de l'objet technique et dans la configuration de l'usager ;*
(4)	2015	*Suivre la trajectoire de l'objet prescripteur* : au fil de sa construction, depuis les premiers tâtonnements des concepteurs jusqu'à sa stabilisation pour une mise en marché, des dimensions politiques et morales se voient inscrites dans le design de l'objet technique ; ce travail itératif d'ajustement des inscriptions se répercute dans la « configuration de l'usager » (Woolgar, 1991) ;
	2005	L'ancrage social et historique des usages dans un ensemble de macrostructures (formations discursives, matrices culturelles, systèmes de rapports sociaux) qui en constituent les formes.
(5)	2015	*Retracer l'ancrage collectif et historique* des usages dans des séries et séquences structurelles (logiques) qui constituent les formes sociohistoriques de l'usage.

Fig18. Évolution du modèle en 5 axes de l'usage

La seconde rédaction (Proulx, 2015) est présentée comme recadrage intégrant les pistes méthodologiques développées par Jérôme Denis (Denis, 2009) en conclusion d'un article présentant l'évolution de la sociologie de l'usage. Les pistes suggérées par J. Denis, inspirées d'Howard Becker (Becker, 2002), seront abordées ultérieurement à l'occasion d'une réflexion sur les méthodologies de l'observation.

L'évolution de la proposition de S. Proulx appelle quelques commentaires.

Tout d'abord, le passage d'une première rédaction à visée structurante à une seconde de nature méthodologique est le reflet d'enjeux qui se sont déplacés. Situer les enjeux contemporains au plan méthodologique n'est pas neutre. Cela témoigne de la difficulté à articuler les méthodes d'observation entre elles, dont celles appuyées sur les traces d'usage. Cela révèle un équilibre nouveau (ou celui de sa recherche) dans un champ d'étude en évolution, privilégiant par nécessité une approche plus empirique et moins théorique.

Ensuite, les deux premiers niveaux (1), (2) renvoient au couple sujet – instrument dans la triade qu'ils forment avec l'objet sur lequel ils opèrent, pour les modèles de description des situations d'activité instrumentée (SAI) (Rabardel, Verillon, 1985, p52). Ces deux niveaux se justifient dans le cadre de cette description. Ils sont, par ailleurs, le reflet historique d'une démarche d'ouverture progressive du champ d'étude, introduisant des problématiques de plus en plus abstraites et nécessitant une compréhension de plus en plus fine et interdisciplinaire. C'est ainsi que, partant de l'interaction contrainte par l'instrument (dominante technique), s'envisagera la prise en compte de

la réalité socio-technique de ce dernier et que suivra l'élargissement des approches aux logiques de l'instrumentation de l'activité (dominante sociale). Ce dernier point se positionne dans un espace recouvrant les niveaux (2) et (3) du schéma. Cet effet de « tuilage » n'est pas incompatible avec l'ambition de la schématisation, il traduit la difficulté à étalonner les échelles et à borner les domaines (évoqué également par l'auteur).

La référence *Human-Computer Interaction* (HCI[117]) qui figure uniquement sur le premier niveau du schéma est équivoque. Elle fit rejaillir une difficulté à définir une frontière entre les problématiques d'interfaces et celles de la définition des applicatifs. Ce problème s'est posé dès l'origine des travaux sur les IHM dès lors qu'il est question d'assurer une médiation vers des applicatifs info-communicationnels[118]. Les deux niveaux interface et applicatif se confondent alors aisément dans leur subsidiarité. L'analyse des médiations info-communicationnelles reposant sur des services du Web de même nature nous situent exactement dans ce cas.

D'un point de vue disciplinaire et méthodologique, les apports de l'IHM ne peuvent se réduire au premier niveau d'analyse (1) si nous suivons la définition hiérarchique qu'en donne l'auteur[119]. Dans son évolution, l'IHM intègre les trois premiers niveaux, c'est-à-dire, le couple usager-interface, l'analyse fonctionnelle et l'organisation de l'activité (tâches), l'analyse située des pratiques. En revanche, les finalités diffèrent effectivement entre l'IHM et la sociologie de l'usage, principalement dans la compréhension du niveau (2). Pour l'IHM, la coordination concepteur / usager est liée à la description normative des tâches. Les mécanismes d'ajustement aux situations atypiques et non maîtrisées ressortent des problématiques de la résolution de problèmes.

En revanche, la description proposée en référence aux travaux de Madeleine Akrich met l'accent sur l'adaptation fonctionnelle des concepteurs. Les spécificités et les approches disciplinaires nécessitent dans le cas présent des ajustements qui méritent d'être soulignés et approfondis.

Une autre lecture de ce tableau consiste à distinguer chacun des axes comme des espaces problématiques indépendants, justifiant ainsi l'association de l'IHM uniquement sur le premier axe, mais cette hypothèse ne nous parait pas plausible.

D'un point de vue méthodologique, la référence à l'IHM se justifie par son apport à l'instrumentation du poste de travail et à la mise en œuvre expérimentale d'environnement pour l'évaluation des pratiques effectives (chapitre 1 et chapitre 3). La prescription de suivi de l'utilisateur va dans ce sens et mobilise implicitement les traces d'usage auxquelles on ne peut échapper (au moins au premier niveau). Pour le second niveau, la référence aux travaux de M. Akrich est une manière d'établir un pont avec le modèle de la traduction que S. Proulx encourage de longue date (Proulx, 2001). L'apport de la catégorie sociologique de la traduction est de saisir l'instabilité des

---

[117] Nous traduisons par IHM (cf. chapitre 1).

[118] Ce fut notamment le cas avec les applicatifs info-documentaires dont la définition a suscité en son temps de vives discussions entre les SIC et l'informatique

[119] La hiérarchisation est décrite par l'auteur dans les termes suivants : « *La nouvelle topique se présente davantage comme un méta-modèle qui met en relief et hiérarchise les principaux niveaux d'analyse pouvant être mobilisés selon différents angles de vue sur les pratiques et les situations d'usage.*». Nous concevons cette hiérarchisation comme une "profondeur de champ", plus l'indice est élevé plus le champ est large.

objets techniques et la complexité des réseaux socio-économiques œuvrant à leur définition. Saisir ces réseaux dans leur dynamique est du point de vue de l'auteur une nécessité méthodologique pour comprendre l'offre technologique et l'usage qui en découlent. L'injonction de suivre le *cours d'actions de coordination* (2) témoigne de ce rapprochement. Les méthodes sous-jacentes à la conduite du suivi n'apparaissent pas spécifiquement liées au courant de la sociologie de la traduction. Dans l'article de 2001, Proulx donne quelques indications méthodologiques qui relèvent davantage de pistes possibles que de méthodes précises. Toutefois, l'indication de suivi mobilise des méthodes susceptibles de rendre compte ou de s'appuyer sur les évolutions spatiales et temporelles des objets suivis.

Dépasser le cadre de la compréhension SAI des objets ou artefacts constituent l'enjeu que traduisent les trois derniers axes (3-5). L'axe (3) correspond à l'inscription sociale des pratiques instrumentées. Le focus est très clairement le social qui se donne à comprendre dans l'inscription des techniques. Les axes (4) et (5) jouent un rôle beaucoup plus *méta* dans la description. Leur interprétation se prête bien à la définition du dispositif telle que nous l'avons envisagé au chapitre 1. L'inscription de dimensions politiques et morales pour l'axe (4), l'historicité de l'axe (5) rejoignent les enjeux de cadrage et de pouvoir portés par la notion de configuration que nous avons évoquée.

Pour ce qui concerne la méthodologie, les deux niveaux (3) et (5) se caractérisent par une approche descriptive. Ceci laisse penser que l'analyse relève d'un niveau supérieur défini (par exemple (4) pour (3)), d'une autre discipline ou d'un autre champ d'étude. Ce qui serait le cas pour (5) qui rappelle les problématiques de l'ethnotechnologie.

3. Conclusions

Les différentes approches méthodologiques que nous avons proposées laissent une impression de découplage entre les réflexions se rapportant aux traces d'usage et les différents courants (ou champ d'études) s'intéressant à l'analyse des pratiques effectives. Ce découplage s'interprète de plusieurs manières qui renvoient à des enjeux disciplinaires et épistémologiques différents.

Il est indéniable que les SHS et les SIC en particulier témoignent, à l'encontre des techniques des traces d'usage et des méthodologies d'analyse qui s'y rapportent, un sentiment ambivalent et fortement polarisant.

D'un côté, il se traduit par de nombreux travaux motivés par la nouveauté technologique mais qui se heurtent aux difficultés d'une appropriation disciplinaire. Les difficultés sont celles de techniques et de méthodes pour le traitement des données (informatiques et statistiques essentiellement) réinterprétées ou difficilement assimilées à partir des propositions méthodologiques du domaine de la fouille des données d'usage.

De l'autre, il se traduit par une mise à distance prudente relevant de la posture disciplinaire plus ou moins argumentée en raison de spécificités méthodologiques ou théoriques.

En cela, il n'y a rien de vraiment nouveau dans cette polarisation des débats scientifiques et disciplinaires. Les conditions sont celles de l'innovation socio-technique qui affecte l'espace social

des sciences comme l'ensemble de la société. Dans le cas des SIC, ces difficultés entrent en résonnance avec le fondement même de la discipline (Ibekwe-Sanjuan, 2014).

Les réserves formulées à l'encontre des traces d'usage et de ses techniques dépassent le cadre de ce débat récurrent. Elles se développent sur un substrat complexe, à la fois idéologique et éthique, qui densifie et légitime la position adoptée. Si ces questions sont effectivement légitimes et importantes, non seulement pour les pratiques scientifiques mais aussi pour le projet de société, il nous parait indispensable et urgent d'investir ces techniques et de les prendre en charge dans une réflexion méthodologique. La puissance de la recherche technologique, débordant largement les sphères académiques, doit nous prémunir du syndrome et de la critique de l'apprenti sorcier. S'emparer de ces sujets n'est pas se soumettre à l'injonction ou aux attentes du marché et de l'industrie de la donnée mais c'est, au contraire, opposer aux pratiques qui tendent à se généraliser les moyens d'une analyse critique.

CHAPITRE 3
Traces d'usage

Toute connaissance acquise sur la connaissance
devient un moyen de connaissance
éclairant la connaissance qui a permis de l'acquérir
(Morin, 1986, p232)

Le titre de ce chapitre : *traces d'usage* renvoie à des articles publiés en anthropologie et en préhistoire, se rapportant aux caractéristiques de l'usure de la surface des outils en silex du néolithique[120].

La recherche en préhistoire se rapporte à des temps antérieurs à l'utilisation de l'écriture. Elle accorde, de ce fait, une large place aux objets qui conservent la marque involontaire du passage des hommes ou l'inscription d'une intentionnalité laissée par nos lointains ancêtres. L'analyse des traces, c'est-à-dire des objets eux-mêmes, naturels ou manufacturés (artefacts) ainsi que leurs altérations, constitue la principale (si ce n'est l'unique) entrée méthodologique à disposition des chercheurs[121].

Les silex taillés sont l'expression d'une connaissance technique qui s'est affinée progressivement et qu'ils documentent. Suivant la nature des polis et des stries observés au microscope sur des silex taillés, les spécialistes du domaine de la tracéologie ont développé des techniques d'analyse qui leur permettent d'en identifier les causes, à savoir : la matière travaillée et le geste accompli. Ainsi, en remontant la chaîne de causalités, il est possible d'inférer une finalité instrumentale à un silex, par exemple celle de faucille. Cette technique d'analyse ayant été affinée expérimentalement, il est désormais possible de prendre en considération des facteurs tels que le degré d'humidité de la matière travaillée et ainsi de suggérer par exemple des usages de l'eau (tannerie) ou des considérations sur le calendrier des moissons (blé vert). Autrement dit et suivant les réserves méthodologiques qui s'imposent et que rappellent les auteurs, l'étude des stries et polis fait émerger des caractéristiques de traces assimilables à un langage à partir duquel la lecture de la surface tranchante des silex est possible. Cette compréhension dépasse la caractérisation de l'outil. Elle permet de formuler des hypothèses concernant des faits et des éléments relatifs aux techniques, aux modes de vie du néolithique.

Nous ne pouvons qu'être admiratifs devant l'ingéniosité des méthodes d'investigation détournées que ces limitations imposent. Nous pouvons tout autant être rassurés et encouragés du fait que des traces d'apparence insignifiante puissent, au contraire, être opportunément signifiantes et productrices de connaissances.

Ce cas extrême de l'analyse des activités humaines fixe une borne méthodologique inférieure à ce qu'il est permis d'envisager dans l'analyse écologique des pratiques associées aux dispositifs info-

[120] Nous prendrons en référence l'article suivant : J. Gysels et D. Cahen se rapportant aux silex (Gysels, Cahen, 1982).

[121] La *tracéologie* s'est d'ailleurs constituée en discipline de l'archéologie préhistorique au début des années 1960.

communicationnels numériques. La coupure avec les conditions de production de la trace est totale. Les réalités humaines sont par trop décalées pour des interprétations subjectives (ou empathiques). Le choix de cet exemple, n'est donc pas uniquement un clin d'œil à une modernité en 2.0 ou un éloge de la sérendipité. On trouve de nombreuses similitudes entre les démarches empiriques.

Il nous permet d'expliciter, en première approche, l'élaboration du concept de *trace d'usage* spontanée, c'est-à-dire non réglée, dans un contexte scientifique strictement constructiviste. Cette élaboration est à deux niveaux : celui de la méthode qui accrédite les hypothèses et celui de la notion qui les supportent.

Dans le cas présent, la trace est la conséquence mécanique directe de séquences gestuelles instrumentées. L'usure observable ne peut pas être imputée au vieillissement naturel, autonome de l'objet. Pour reprendre les termes d'Alexandre Serres, la trace est une marque du passé qui s'interprète au présent (Serres, 2002). L'inscription sur le support est un indice, une trace d'usage. Elle traduit les qualités intrinsèques de l'artefact (dureté relative de la pierre) et la logique de son usage (trancher) rapportés à son histoire (indéterminée). La trace archéologique est ainsi une réduction de l'information originelle. En première approche, l'usage est à concevoir dans un sens restreint ou faible d'utilisation. Le terme d'utilisation n'appelle ni la régularité ni la monotonie que l'on fait porter au sens fort du concept d'usage (Perriault, 1989).

La démarche scientifique qui permet l'attribution du sens fort est le résultat d'une élaboration reproduisant les trois temps de la méthode scientifique de C. Peirce citée par Katia Angué (Angué, 2009).

Premier temps.

> Pour que la trace acquière le statut d'indice, il est nécessaire que les éléments qui la caractérisent perdent dans l'observation leur caractère accidentel. Pour cela, il faut que le silex et ses traces puissent être rapprochés d'une collection constituée parmi des silex présentant des stigmates d'usages similaires. La définition d'une collection traduit la reconnaissance d'un ensemble de caractéristiques communes spécifiques de celle-ci, à partir desquelles se construit la fonction de similarité qui permet de la constituer effectivement.

Second temps.

> Ce temps est celui de la formulation d'une hypothèse et de l'expérimentation analogique qui permet d'établir une chaîne de causalité (hypothético-déductive) qui, partant du réel (l'usure polie / strie) formule le procédé et le matériau qui les produisent. Une fois établie, cette règle productive fait l'objet d'une validation empirique à partir d'un échantillon représentatif.

Troisième temps.

> L'interprétation de la nature de la collection vient alors de la similarité observée entre chacun des éléments de l'échantillon du test avec ceux de toute la collection. Sous cette condition, on peut raisonnablement inférer (induction) que l'hypothèse empirique, à savoir la répétition d'une séquence gestuelle spécialisée, est plausible en tant qu'explication (prémisse) de la collection.

À la suite, la trace observée devient distinctive d'usage technique (coupe d'herbe) et prend valeur d'empreinte pour une classe d'activité dont l'identification (moisson) est un enjeu plus fort nécessitant le croisement de faits.

Ce qui ressort de l'analyse de ce cas et qui peut se transposer à nos travaux, c'est le régime de production de connaissances qui est à l'œuvre. Il est ici fondamentalement de nature abductive, c'est-à-dire qu'il fait porter sur la formulation des hypothèses se rapportant aux logiques d'usages (A->B) le potentiel d'interprétation des traces d'observation (B). C'est dans ce sens que les traces d'usage vont nous intéresser comme manifestation et support méthodologique pour l'exploration des pratiques instrumentées (A).

Plaçons-nous désormais à l'ère contemporaine où de nouvelles perspectives sont ouvertes, transformant tout objet ou logiciel en objet connecté (*smart system*), communiquant et interagissant dans le monde réel ou dans la virtualité[122] d'univers numériques. La spécialisation et la finalité dont étaient porteur l'artefact, qui l'identifiaient et le qualifiaient pour l'agir instrumenté tendent à disparaitre. L'outil devient générique, apte à de multiples réalisations. Il devient alors très difficile d'identifier l'activité accomplie par son instrumentation, l'usager orchestrant celle-ci à la demande. Cette déspécialisation affecte également les modalités de l'interaction. Après la main, organe d'expression majeur, le corps dans son ensemble est investi de la fonction instrumentale[123].

L'environnement connecté dans lequel nous évoluons s'investit de potentialités instrumentales permanentes et ubiquitaires. L'accessibilité permanente aux fonctions de localisation, la généralisation des espaces couverts par un accès wifi, sont deux illustrations de cette tendance temporelle et spatiale. La stabilité en qualité de l'offre technique assure une standardisation des conditions d'usage, à partir de laquelle s'élabore une offre de plus en plus complexe de services et de moyens techniques. Nous pouvons considérer l'effet *pervasif* des technologies numériques comme un effet rétroactif, politique et normatif des logiques de dispositifs socio-techniques visant à sa préservation et à son extension. Ainsi, ces potentialités permettent aux dispositifs info-communicationnels, par des ajustements locaux (disponibilité, etc.), de se déployer et de se configurer pour assurer la conduite de l'activité, légitimant à leur tour les conditions de leur émergence.

La permanence de l'effectivité des objets connectés (smartphones, etc.), l'étendue spatiale de leur plage d'utilisation ont été évoqués dès l'avènement de la 3G comme appelant un renouvellement méthodologique pour en analyser les effets sociétaux. Comme nous le verrons, celui-ci a été entrepris mais il ne s'est pas affranchi des objets. Le fil directeur que nous tiendrons dans ce chapitre est celui correspondant aux situations de *smart system* ou d'objets connectés. À notre sens, il s'agit d'une nécessité pour assurer une plus grande portée à la réflexion méthodologique se rapportant à l'observation des pratiques effectives.

Les objets connectés - qui informent, s'informent et communiquent en permanence, élaborant et partageant des représentations, consignant la contribution active ou passive, volontaire ou

[122] Virtualité qui n'est pas systématiquement associée à une coupure du lien social.

[123] Nous utiliserons la définition du geste instrumental décrit par Claude Cadoz (Cadoz, 1994). cf. chapitre 4, §1

involontaire des individus sous leur portée - deviennent en eux-mêmes des supports méthodologiques incontournables. C'est dans ce contexte qu'il nous faut définir le concept de trace numérique d'usage.

1. Constitution d'un objet numérique de suivi

1.1. La trace numérique

Une trace numérique est en premier lieu une représentation numérique informatisée, c'est-à-dire une séquence délimitée d'octets inscrits dans une mémoire éphémère ou durable. Pour constituer une trace numérique, la séquence délimitée doit avoir une unité qui tient en premier lieu à la position occupée en mémoire. Celle-ci lui attribue le sens de codage numérique ou de donnée et la distingue des programmes (autre sorte de données).

Dans le domaine informatique, il est très peu probable d'être confronté à des traces numériques dont on ignorerait tout. Toute trace numérique est cohérente par rapport à un état du système de traitements qui l'a produite. Cette simple marque de fabrique permet d'établir que sous une apparence uniforme (bitstream), existent un encodage, une syntaxe et une sémantique la rapprochant d'une donnée semi-structurée à un instant donné du traitement. Cette définition est la plus générale que nous puissions formuler de la trace numérique. Elle met en avant l'une de ses caractéristiques ontologiques : la trace numérique n'est pas neutre. Elle est toujours porteuse d'une intentionnalité qui n'est pas nécessairement associée à une finalité opérative. En revanche, dès lors qu'une trace numérique est produite, elle est susceptible d'un traitement informatique.

Nous définirons alors comme *trace numérique d'usage*, l'enregistrement en mémoire secondaire d'une trace numérique constituée durant le fonctionnement du dispositif info-communicationnel. Cet enregistrement est soit le résultat d'une action volontaire de l'un au moins des acteurs du dispositif, soit il découle de la configuration technique et des traitements qui s'y déroulent. Dans cette première définition, l'usage est considéré au sens faible d'utilisation et non celui d'une pratique régulière.

1.2. Intentionnalités des dispositifs info-communicationnels

Pour rester cohérent avec le concept de dispositif, nous établirons une distinction se rapportant à l'*intentionnalité* (Searle, 1985) des acteurs engagés dans le dispositif[124]. L'intentionnalité gouverne une partie des actions que chaque individu conduit. L'intentionnalité nous permet de nous référer au cadre mental du projet d'action plutôt qu'aux séquences d'actions supposées conséquentes que l'observateur ne peut pas connaître dans le détail si ce n'est dans la trace résultante qu'elle aura produite.

[124] L'intentionnalité est une notion qui a fait l'objet d'un grand intérêt dans les travaux se rapportant au dialogue Homme-Machine. Inspirée des travaux de H. Grice et de J. Searle, l'intention (*intent*) intervient dans de nombreux modèles de planification des actions en complément de la notion de croyance (*belief*). Ces notions sont utilisées comme support de la pragmatique de l'énonciation (actes de langage).

Nous distinguerons une intentionnalité première, qui s'exprime et s'actualise dans le fil des interactions, d'une intentionnalité seconde inscrite dans l'histoire du dispositif mais dont on a perdu la justification et qui reste active alors même qu'elle est atténuée au présent. Dans cette perspective, toute trace numérique est le résultat combiné d'intentionnalités premières et secondes.

Chacun des objets connectés est porteur d'une intentionnalité seconde, elle-même portée par l'ingénierie nécessaire au fonctionnement du système informatique dont est issue la trace numérique. Cette ingénierie est opérée par différents acteurs intervenant dans les différentes phases de :

- l'élaboration des éléments logiciels constitutifs du système informatique : les contextes peuvent être ceux d'éditeurs indépendants, de développement spécifiques externalisés ou bien internes à l'organisation administrant le système ;
- l'installation des différents composants techniques, matériels ou logiciels qui caractérisent le système informatique depuis le système d'exploitation jusqu'aux logiciels applicatifs. Ici, doivent également être pris en compte les mises à jours de différente nature, effectives ou non, les patchs réalisés pour combler des lacunes ou assurer la cohésion du système, les bricolages techniques de maintenance, etc. ;
- le paramétrage des différents éléments, matériels ou logiciels qui composent le système informatique. Celui-ci peut être réalisé soit par défaut, soit calibré sur des indicateurs fixés par les constructeurs ou les éditeurs ; ils peuvent aussi être adaptés localement, en réponse ou non à des contraintes de fonctionnement ou des objectifs spécifiques.

L'intentionnalité de second niveau est, pour la plupart des systèmes, faiblement coordonnée pour ce qui concerne les traces numériques d'usage.

La trace d'usage est porteuse de problématiques fonctionnelles et normatives. Ce qu'elle enregistre doit permettre d'établir, à un niveau de traitement donné, une mesure d'écart à la normalité d'un fonctionnement ou d'une utilisation prévue. La nature des contenus et la durée de leur pertinence n'est pas garantie en dehors des cycles de la surveillance fonctionnelle. L'absence de coordination est *a priori* préjudiciable à la qualité informationnelle. En particulier, l'historisation conjointe des traces et des cadres interprétatifs qui leur sont associés n'est pas assurée. Dans ce contexte, il est très difficile d'aller vers une représentation et une caractérisation individuelle. De manière concrète cela signifie que le plus souvent les fichiers de *logs* sont inutilisables sans un gros travail de réingénierie informatique préparatoire pour les rendre opérationnels par la suite. Une fois exploitable, les traces contenues dans ces *logs* doivent encore être articulées.

Il est cependant des cas où, au contraire, il existe une intentionnalité maîtresse, fondamentalement première, qui guide un projet très structurant de prise en charge d'une trace numérique d'usage dans l'élaboration du système informatique. Dans ce contexte, la structuration informationnelle de la trace est maîtrisée dans les différentes strates de traitements et contribue ainsi à l'unité interprétative de celle-ci. Cette unité est de nature sémantique (structuration et contenus) mais aussi pragmatique par le suivi complémentaire des conditions de production de la trace. Elle se traduira dans ce cas par la mise en œuvre de composants spécialisés voire de l'inscription de la trace

numérique d'usage dans le système d'information (SI). Dans ce contexte, la trace d'utilisation devient véritablement dans son historisation, la trace d'un usage singulier.

2. Enjeux de la traçabilité systématique

Les cas dans lesquels se manifeste une intentionnalité coordinatrice se rencontrent lorsque les finalités du système informatique sont régies par un jeu de contraintes entre les logiques d'exploitation et les normes règlementaires.

2.1. L'*accountability* : de la sûreté de fonctionnement, à la responsabilité éditoriale

Le principe de traçabilité des actions des usagers est historiquement associé aux problématiques de la sécurité informatique et cela bien avant l'avènement de l'Internet grand public.

Très tôt dans l'histoire de l'informatique en France, les centres informatiques ont eu comme obligation légale d'avoir en permanence une imprimante dont la fonction était expressément de produire le *listing* de l'ensemble des opérations exécutées sur le système d'information. Le document imprimé, désigné comme *la trace,* était conservé constituant une archive à valeur juridique.

L'axe principal de la sécurité informatique concerne la sûreté de fonctionnement, c'est-à-dire la garantie de la mise en œuvre du service informationnel. L'accent est mis sur les trois attributs clefs que sont : la confidentialité (impossibilité d'une divulgation non autorisée) l'intégrité (impossibilité d'une altération inappropriée) et enfin la disponibilité du service d'information lui-même. À ces attributs qui portent aussi sur les fonctionnalités communicationnelles du système s'en ajoutent d'autres, secondaires, parmi lesquels deux se rapportent à la personne :

- la non répudiation (*nonrepudiability*) renvoie à l'identité des individus émetteur/récepteur impliqués dans la circulation d'un message. Celle-ci doit être établie avec certitude et conserver son intégrité ;
- la responsabilité (*accountability*) renvoie à la conduite d'opérations individuelles dans le système informatique, suivant des critères analogues.

La traçabilité se rapporte à ces deux attributs. Elle implique un principe d'authentification et de signature (identité numérique) de tout acte impliquant une personne qu'elle ait un rôle d'opérateur direct ou indirect (récepteur). Une trace d'usage devient par extension, l'association d'une représentation d'un acte rapportée à une signature au sein du système informatique. Elle est rendue opérationnelle par la journalisation ou *logs* (enregistrement horodaté) de l'ensemble des traces d'usage. Cette représentation événementielle de la trace d'usage est cohérente avec la mise en œuvre des traces d'exécutions (*logs* systèmes) que gèrent les administrateurs de systèmes informatiques pour en contrôler le bon fonctionnement.

Le développement concomitant des systèmes d'information et de l'Internet a étendu le sens de l'*accountability*. De nouvelles normes et réglementations ont vu le jour, amenant un accent plus marqué sur la sécurité de l'information (normes ISO/IEC de la série 27000) et sur la circulation de documents.

Dans le cadre de la loi pour la confiance dans l'économie numérique (LCEN) promulguée le 21 juin 2004 et mise en application par décrets progressifs, la traçabilité des personnes et de leurs actions est une obligation pour les prestataires publics ou privés qui offrent des services permettant aux internautes la publicisation de contenus (décret 1er mars 2011). En conséquence, toute offre de service d'accès à l'Internet implique de la part de l'offreur le suivi et la journalisation (*log file*) des sessions authentifiées et l'obligation de tracer les matériels et les actions individuelles sur des contenus. L'enregistrement ainsi réalisé doit être conservé pour une durée légale d'un an[125]. Ainsi, à partir d'une origine strictement technique, la trace numérique d'usage s'est institutionnalisée en règle de contrôle individuel. Elle porte en elle un double impératif : la transparence de l'action et la responsabilité, que résume l'anglicisme d'*accountability* associé aux discours sur la moralisation de l'action publique et politique et la justification de projets de type *open data*.

2.2. Perspectives analytiques et limites de la journalisation Web

Dans cette partie nous ne nous intéresserons qu'aux éléments propres aux dispositifs dans une exploitation ordinaire de ceux-ci. Nous reviendrons sur les adaptations possibles des contextes expérimentaux plus tard.

Le développement du Web grand public a mis en évidence l'importance de la journalisation produite par les serveurs applicatifs. Au-delà des logiques de maintenance et d'administration des services, l'enregistrement des traces d'usage a ouvert, au début des années 2000, de nouvelles perspectives stratégiques pour les promoteurs de sites Web.

Les *logs* des serveurs HTTP qui assurent la circulation des contenus HTML sur le Web sont de ce point de vue les plus significatifs de ces enjeux. C'est à ce niveau applicatif que s'organisent les interfaces utilisateur structurant des parcours documentaires et l'accès aux services proposés. Les traces d'usage enregistrées dans ces journaux rendent compte des interactions protocolaires entre le navigateur (client) et le serveur ainsi que des flux de données et de documents que ces interactions canalisent. En fonction du paramétrage du serveur HTTP, ces données de *logs* constituent une ressource pour l'évaluation du degré (volume, intensité) et des modalités d'utilisation du site ainsi que des contenus ou des services hébergés.

Les *proxies* Web qui font partie de l'écosystème du protocole HTTP sont susceptibles également de comporter des fichiers de journalisation. Outre la difficulté de mise en œuvre et les effets de bords sur la performance de ce genre de serveur, les ficher de *logs* de *proxies* n'offrent qu'une vue localisée ne permettant pas d'obtenir une vue de l'ensemble des requêtes soumises aux serveurs. En revanche, la mise en œuvre d'un proxy peut être intéressante dans un contexte expérimental se rapportant à une communauté identifiée et des dispositifs paramétrables.

Les traces d'usage récoltées côté serveur orientent vers deux types d'analyses centrées respectivement sur le site (*site centric*) ou sur les usagers (*user centric*). Ces types d'analyses ont été envisagés simultanément. Cependant, c'est l'analyse centrée sur les sites qui est la plus aboutie. Cela s'explique aisément du fait que les traces s'interprètent immédiatement comme des parcours. En

[125] Décret n°2011-219 du 25 février 2011. Au-delà les données doivent être anonymisées.

revanche, l'interprétation de la trace individuelle nécessite aussi des modèles interprétatifs du processus de parcours qui sont à établir dans le contexte du site.

2.2.1 Les limites techniques

Les possibilités d'analyse basées sur les données disponibles sur le serveur restent limitées.

En effet, les protocoles de communication sur Internet, les adaptations techniques comme les *proxies*, l'attribution des IP à la volée (DHCP)[126] ou encore les mémorisations *cached pages* (sur le navigateur ou les *proxies*) dégradent la qualité de l'information référentielle transmise au serveur. On peut ajouter à cela, le balayage automatique des sites par les robots qui interfère avec l'enregistrement des traces d'utilisateurs.

Enfin, l'horodatage est également source de difficultés puisque l'information temporelle est établie par le serveur à la réception des requêtes. Les délais d'acheminement et de traitements qui émaillent l'échange d'information entre le client et le serveur ne peuvent pas être appréhendés facilement. En outre, l'absence d'information signifiant l'utilisation de pages cachées conduit à des erreurs d'interprétation sur les temps de lecture.

Des solutions complémentaires, côté clients[127] se sont développées pour pallier ces imprécisions, sans pour autant tout résoudre. Elles reposent sur deux techniques éventuellement conjointes :

- le *cookie* qui est aussi une trace numérique du protocole HTTP pris en charge par le navigateur, dont la fonction est d'assurer une mémorisation locale sur la machine client d'informations de session.
- le *tag* qui désigne un segment java script inscrit dans le corps de la page, dont l'exécution dans le navigateur du client produira un signal de visite auprès du site d'enregistrement ;

Ces différentes techniques nécessitent de maîtriser une ingénierie informatique pour assurer une remontée d'information précise sur l'origine de la requête au serveur. Cela suppose également une ingénierie informationnelle pour redéfinir la trace d'usage articulant les données du cookie avec celles du log. La résolution technique n'est pas la seule difficulté, il faut aussi que la méthode soit conforme aux réglementations en vigueur[128] (obligation du recueil du consentement, etc.) et qu'une relation de confiance soit établie avec les usagers.

Pour ces différentes raisons, la mise en œuvre des techniques d'enrichissement et d'analyse des données d'usage est à l'origine d'offres de services qui composent le marché de l'analytique du Web

---

[126] *Dynamic Host Configuration Protocol*

[127] Au sens client-serveur, c'est-à-dire du côté du poste de l'usager.

[128] Article 5 de la directive du 12 juillet 2002 sur la protection de la vie privée dans le secteur des communications électroniques. Voir également article 29 sur la protection des données du 22 juin 2010. http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2010/wp171_fr.pdf

(*Web analytics*). Ces solutions externalisées auprès de tiers spécialisés, sont par nature, plutôt « centré site » pour les contractants[129] bien que des solutions hybrides se dessinent à l'horizon[130].

2.2.2 Approche centrée site

Pour l'approche centrée site, les méthodes d'analyse des données extraites des *logs* du Web ont rapidement fait l'objet d'outils qui se sont eux-mêmes bien intégrés aux logiques du marketing et de la communication marchande.

Les éléments quantitatifs mais aussi qualitatifs que ces outils permettent de dégager, contribuent à une rationalisation de l'évaluation de la performance info-communicationnelle des contenus et des services en ligne. Ils conduisent à une approche itérative et empirique de la conception du site fondée sur l'intégration des comportements effectifs observés.

Ils permettent également, moyennant une imprécision que l'on ignore souvent, de tester des hypothèses se rapportant aux contenus ou aux services en ligne et de les évaluer rapidement. Cette dynamique d'ajustement va dans le sens d'une plus grande réactivité aux mouvements de l'opinion (audience) comme de ceux du marché.

La pression du temps caractéristique du développement actuel des marchés se nourrit de cette mise en tension, légitimant en retour l'instrumentation au sens large (métriques et procédés) de la trace d'usage. Pour les acteurs du secteur de la communication commerciale, cette possibilité de mesure est l'une des origines d'une culture de la performance qui s'est propagée à l'ensemble de l'économie du Web (Ref.13). On constate le même effet de contagion *métrologique* dans le secteur des médias grand public, renforçant ainsi le rôle des acteurs de l'analyse d'audience et d'opinion (Ref.18).

2.2.3 Approche centrée usager

Pour l'approche centrée usager, les objectifs poursuivis sont multiples. Ils peuvent consister dans la mise en œuvre de services d'accompagnement personnalisé de l'usager ou de services d'enrichissement de représentations à finalité stratégique pour le système. Parmi ces services à haute valeur ajoutée, se distinguent les modèles comportementaux qui conduisent à la proposition de profils caractéristiques de classes d'usagers voire à la caractérisation des représentations personnelles. Ces services ont généralement une finalité interne mais sont aussi une source de revenus potentiels associée au marché de la monétisation des bases d'usagers (Ref.13).

Ces objectifs sont cependant tributaires de la connaissance dont on dispose originellement sur chaque usager. En effet, celle-ci peut être réduite à une adresse IP c'est-à-dire à une identité *a priori* collective, compilant l'ensemble des individus qu'elle connecte au Web. Au contraire, elle peut être en relation avec un identifiant interne relié à un système de gestion d'abonnés.

---

[129] Elle peut être « user centric » pour les prestataires de ces services qui croisent une énorme quantité de données de ce type.

[130] La solution *Universal Analytics* proposée par Google (2013) témoigne de cette évolution en cours qui s'intègre avec le système d'information et qui permet d'envisager des analyses beaucoup plus fines et adaptées aux logiques de chacune des entreprises.

Sans connaissance externe au journal, les possibilités d'investigation sont très réduites. Si l'adresse IP est la seule information caractéristique, on peut seulement espérer faire émerger, mais avec une grande incertitude, des classes de comportements susceptibles de guider une analyse très globale des populations auxquelles la ressource s'adresse.

Pour aller plus loin, il est nécessaire de disposer d'une connaissance plus fine de l'usager. Nous avons déjà souligné que des moyens complémentaires comme les cookies peuvent être employés. Des situations intermédiaires existent. Elles s'appuient sur la définition de panels d'usagers dont les environnements sont équipés de sondes (*meter*) ; ces panels définissent des données de référence sur lesquelles s'agrègent les données d'individus anonymes. La méthodologie est empruntée aux méthodes des instituts de sondage et d'analyse de l'audience.

Si l'accès au site est authentifié, l'identifiant[131] de l'individu permet une agrégation plus sûre des traces et de là, une analyse en classes de comportements plus juste. Une analyse nominative est aussi possible mais elle ne fait de sens qu'en liaison avec un système de gestion d'abonnés relevant d'une logique d'activité[132]. À ce stade, il convient de s'interroger sur le sens que revêtent les données de journaux du Web par rapport aux données du système d'information qui, par hypothèse, est étroitement couplé aux processus métiers. Les processus métiers décrivent en l'occurrence, dans les termes du système d'information, une formalisation de l'activité développée au sein d'une organisation.

Pour que l'analyse fasse sens, il faut que l'interaction site-usager s'inscrive dans l'un des processus métiers ou qu'elle soit l'occasion d'une captation de valeur. La personnalisation de l'interaction, des contenus ou des services entre dans le cas des processus métiers. Ceux-ci peuvent aller très loin dans la connaissance de l'individu. Le développement des services égocentrés, dont font partie les réseaux sociaux, ont par nature et légitimité vocation à apprendre de l'abonné et à restituer dans un cercle de publicisation plus ou moins étendu une représentation qui en découle.

La captation de la valeur est pour ce qui précède, un méta processus à partir duquel s'envisage la régulation ou la réorganisation des processus métiers. La prédictibilité des comportements de l'usager en fait partie. Elle constitue un enjeu extrêmement fort et incertain pour le management stratégique qu'il soit lié à l'opinion ou au marché. Elle rejoint des enjeux de recherche qu'il convient de situer.

3. La fouille des données d'usage du Web

Le *Web Mining* est le rapprochement opéré entre les techniques du *Data Mining* et les problématiques émergeant de l'ouverture du Web au grand public et à son appropriation stratégique, marketing et commerciale.

La paternité de l'expression *Web Mining* semble devoir revenir à Oren Etzioni dans un article ayant fait date et dont l'intitulé était : *The World Wide Web: cachemire or goals mine?* (Etzioni, 1996). La

[131] Qui peut aussi être partagé…

[132] Ce qui peut être le cadre de la gestion de la relation client (GRC ou CRM) dans un système marchand.

métaphore évoque le processus de recherche, intuitif par des égards, qui conduit à la découverte d'un filon que l'on exploitera ou au surgissement de la pierre précieuse. Elle associe également le mineur et le chercheur de pierres précieuses à cet univers. La traduction française, *fouille du Web*, traduit de manière satisfaisante le processus de recherche mais perd définitivement le sens de la valeur associée à l'objet, voire celui du dépassement que l'on peut aussi associer au chercheur d'or.

Le questionnement repris par Oren Etzioni dans son article traduit le scepticisme manifesté (à l'époque) au regard de la possibilité d'exploiter le Web en tant que source de données en raison de la faible structuration de l'information et la très grande dynamique d'évolution. Deux axes se dégagent positivement selon l'auteur. En premier lieu, celui de la découverte de services et de documents mais suivant le paradigme inspiré des méta moteurs requêtant les index des moteurs de recherche. En second lieu, l'extraction d'information mais suivant le paradigme des moissonneurs (*harvest system*) qui opèrent sur des documents semi-structurés. Le programme qu'il trace alors avec clairvoyance fait référence aux outils et aux méthodes de l'apprentissage automatique (*machine learning*) et de la classification automatique, outils indispensables pour produire des généralisations et des regroupements à partir d'objets informationnels incomplets disponibles à grande échelle, c'est-à-dire en suivant les conditions du *Big Data*.

S'agissant de considérer le Web en tant que source de données disponibles pour traitements automatiques, trois axes majeurs se développent. La *fouille de données d'usage du Web* (*Web Usage Mining* ou WUM) (Cooley, Mobasher, Srivastava,1997), (Vellingiri, Pandian, 2011) est l'un de ces axes ; il s'est constitué spécifiquement à partir de la disponibilité des journaux de serveur HTTP comme source de données[133]. Il coexiste avec les deux autres axes que sont : la *fouille des contenus du Web* (*Web Content Mining* ou WCM) et la *fouille de la structure du Web* (*Web Structure Mining* ou WSM) qui exploitent les pages et les liens exprimés dans le format HTML (Borges, Levene, 2000), (Kosala, Blockeel, 2000).

3.1. Caractérisation de la fouille des données d'usage

Dès l'origine, le processus de fouille des données d'usage a été décrit suivant les trois étapes développées ci-après.

3.1.1. Les prétraitements

Les prétraitements constituent la première phase de la préparation des données. L'objectif de ceux-ci est de constituer une ressource originelle de référence, complétée et stable, à partir de laquelle se construit l'extraction puis l'analyse des données. Nous définirons la notion de segment de trace comme correspondant à une ligne du journal, c'est-à-dire à une sollicitation du serveur. Le segment de trace est l'unité à partir de laquelle se reconstruit la trace individuelle. Les prétraitements se déroulent en trois temps.

La *clarification des données* correspond au premier pré-traitement, il reprend la séquence suivante :

[133] On trouve parfois le terme de fouille de journaux (*Log mining*) pour désigner la fouille des données d'usage du Web.

- Identifier les différents fichiers associés à une partition susceptible d'avoir été réalisée durant la journalisation et de les fusionner dans un fichier unique de référence ;
- Incorporer les données complémentaires à chaque segment de trace. C'est par exemple la provenance du lien de navigation conduisant à la sollicitation du serveur qui peut être enregistrée dans d'autres journaux complémentaires (*referrer logs*) ;
- Éliminer les éléments ayant un statut d'erreur, ne concernant pas les ressources étudiées ou plus généralement non pertinents (nature multimédia de l'objet, feuilles de styles, etc.) ;
- Convertir les éléments pertinents, mais n'ayant pas encore un statut d'éléments analysables en tant que tels (élément multimédia, etc.). Cela peut aller jusqu'à mettre en œuvre des techniques de classification ou de clustering afin de constituer des catégories analytiques. Ces techniques peuvent être étendues aux contenus des pages Web (WCM) voire à la structure du site (WSM) associés aux traces d'usage.

L'association des différentes techniques de fouille du Web (WCM, WSM), peut amener à proposer un enrichissement des données d'usage.

Le second temps porte sur deux prétraitements à l'incidence réciproque et de nature heuristique :

- L'adresse IP étant par nature ambiguë[134], l'individuation des segments de trace est nécessaire. Nous désignons sous ce terme les hypothèses[135] et les calculs conduisant à l'attribution d'un identifiant individuel à chacun des segments de trace. Cette attribution, loin d'être évidente, doit prendre en compte la complexité croissante du Web (multiplicité des *proxies*, préservation d'identité, multicanal, multi-agent, etc.) ;
- La segmentation temporelle en session. Une session est une unité d'interaction se déroulant dans une continuité d'activité de l'usager. La segmentation temporelle s'appuie le plus souvent sur un seuil d'inactivité fixé empiriquement au-delà duquel on assume l'hypothèse de discontinuité.

Le troisième temps est celui de l'adaptation du format d'enregistrement afin d'enchaîner l'étape de recherche de *patterns* ou motifs. Cette adaptation peut conduire à la mise en œuvre d'une base de données.

3.1.2. La recherche de *patterns*

Un pattern est un motif qui se dégage d'un ensemble d'éléments présentant un minimum de structuration interne. Dans le cas présent, la recherche de motifs va s'appliquer sur des regroupements se rapportant à un même identifiant/usager ou à un groupe d'individus.

Suivant les cas, une unité inférieure à la session peut être envisagée. Dans ce cas, on parlera de transaction pour ce regroupement qui peut concerner toute ou partie de la session. Pour mettre en

[134] Elle est non irréfutablement et exclusivement associée à l'activité d'un seul individu.

[135] L'unité de session, l'unité du matériel utilisé (navigateur, objet), sont des éléments pris en compte dans l'identification personnelle.

œuvre un découpage en transaction, il faut disposer d'une connaissance préalable (modèle) de l'organisation du site établissant cette unité.

La détection de *patterns* passe par une représentation de la structure des sites en fonction des URL de pages et par la représentation des séquences parcourues. Les représentations favorisant la description de graphes s'imposent naturellement. Il existe plusieurs types de méthodes analytiques pouvant être mises en œuvre suivant les objectifs de l'analyse.

- L'analyse des *tronçons* (*path analysis*) permet de confronter la structure du site avec les parcours effectifs. Dégager quels sont les tronçons les plus/moins empruntés, les plus/moins longs, etc. soutient une réflexion *site centric* déjà évoquée. Ce type d'analyse sert aussi à situer le site dans son environnement (entrée/sortie) et de mieux comprendre les contextes et la nature des navigations ;
- Les méthodes de découverte de *règles d'association* s'appuient sur la détermination de corrélations entre des séquences de transactions. L'exemple le plus significatif est celui de la recommandation pratiquée par les sites marchands qui *poussent* des produits suivant la règle d'association : "ceux qui ont vu/acheté ce produit ont aussi vu/acheté celui-ci" ;
- La découverte de *schémas séquentiels* porte sur la succession des transactions ou des sessions. L'objectif est de pouvoir prédire, compte tenu d'une séquence identifiée, quelle séquence est la plus probable suivant quelle modalité (délai/tronçon). Ce type d'identification de schémas de visite est utilisable dans les stratégies comportementales appliquées sur des groupes homogènes ;
- La découverte de *règles de classifications* permet dans le cas présent d'attribuer une classe à un nouvel item caractérisé (identifiant) à partir des classes établies depuis des caractérisations associées à une population de référence. La caractérisation est établie soit en fonction d'une connaissance individuelle externe soit à partir d'une catégorisation des transactions ou des contenus ;
- La segmentation (*clustering*) permet enfin de regrouper des items selon des caractéristiques communes, en particulier, les identifiants. Les segments ainsi établis permettent d'envisager un ciblage comportemental sur ces groupes homogènes.

3.1.3. L'analyse des *patterns*

Les méthodes décrites précédemment nécessitent une grande habitude et une connaissance de spécialistes. Permettre le transfert de l'analyse à d'autres acteurs, spécialistes d'autres domaines (en SHS par exemple) est un enjeu extrêmement fort. Cette réflexion fait partie du domaine d'étude du *Web Usage Mining* (WUM). Ces attendus se traduisent par des efforts de développement : en techniques de visualisation, langages et de formalismes opérationnels sur les données et enfin en termes d'interface permettant d'organiser un environnement d'analyse.

3.2. Questionnements actuels

Près de 20 ans après les premiers travaux, il est étonnant de constater, dans les publications actuelles, la très faible évolution de la définition méthodologique et technique pour le sujet de la fouille des données d'usage du Web.

Cela signifie en premier lieu que le processus analytique a été rapidement structuré et stabilisé. Une explication de cette stabilité semble résider dans la généralisation du principe de la journalisation applicative sur le Web. Cette généralisation tient largement au développement de services égocentrés fondés sur la personnalisation et l'individualisation de la relation aux plateformes les proposant. Ces services ont en effet toute légitimité à proposer une authentification et à capitaliser sur le profil individuel, des données susceptibles d'entretenir et d'étoffer l'offre de services.

En conséquence, le concept de trace d'usage du Web se généralise à l'ensemble des traces numériques produites par les usagers dans le contexte de services en ligne, de plus en plus fréquemment authentifiés. La relation personnelle mise en scène par les plateformes dans l'abonnement construit les conditions d'une acceptabilité sociale plus grande vis-à-vis des mécanismes de la traçabilité côté client (dont les cookies).

Ainsi, l'intégration de différents services dans une plateforme permet non seulement d'enrichir l'offre et de créer un écosystème cohérent pour l'usager abonné mais encore de renforcer la quantité et la qualité de l'information journalisée.

Dans ce contexte, l'analyse des données d'usage du Web prend tout son sens et gagne en efficacité. L'enrichissement des profils individuels devient une activité importante et une source de profits dans le secteur de la monétisation des données personnelles avec les conséquences que l'on connaît.

En second lieu, cela indique qu'il s'agit davantage d'un champ d'expérimentation pratique, relevant d'une ingénierie de la donnée que d'un champ théorique. Les méthodes évoquées découlent de modèles généralement éprouvés en informatique et en statistique. L'analyse des données d'usage du Web permet d'en évaluer la performance et l'efficacité dans le changement d'échelle qu'impose la mise en situation de production.

Les enjeux associés à la fouille des données d'usage portent sur l'interprétation en contexte des données et des résultats d'analyse. Il s'agit d'une part de développer des outils permettant une manipulation et une lecture compréhensive des représentations numériques et d'autre part, de constituer une expertise permettant de faire le lien avec une compréhension métier (commerce, etc.) des résultats.

Les publications récentes sur le WUM[136] tentent de dégager une orientation pour les travaux à venir. Le plus souvent cela se traduit par l'évocation de cas d'application ou de secteurs prometteurs. Ces hésitations que l'on perçoit à la lecture pourrait correspondre à une phase transitoire au sein d'une réflexion sur un objet d'étude qui est peut être au bout de ce qu'il portait ou qui nécessite d'être redéfini pour permettre de franchir une nouvelle étape.

---

[136] *Web Usage Mining*

4. Conclusion.

Dans l'univers numérique, la trace d'usage est une élaboration, résultat d'une intentionnalité plus ou moins coordonnée d'acteurs du dispositif. L'intérêt pour les traces d'usage s'est constitué progressivement dans la continuité du développement des méthodes de sûreté des systèmes informatiques. Les techniques de journalisation ont ainsi gagné la mise en œuvre des services sur le Web. Suivant le degré de coordination, il est possible à partir de ces traces, par une analyse longitudinale, de dépasser l'information ponctuelle d'une utilisation anonyme propre à répondre aux besoins d'une approche strictement centrée site. Pour cela, les journaux des serveurs Web ne sont, le plus souvent, pas suffisants et doivent être complétés par des informations exogènes se rapportant aux usagers.

La méthodologie d'analyse que requièrent ces données authentifiées s'est constituée en un domaine d'étude (*Web Usage Mining* ou WUM) dont l'intérêt a été très vite perçu par les acteurs de l'économie numérique. La mise en œuvre de plateformes de services égocentrés a favorisé leur développement, suivant une finalité de capture de la valeur associée à la connaissance fine des personnes, ouvrant ainsi la voie au ciblage comportemental et à la monétisation des données personnelles.

Comme nous l'avons souligné, les techniques du WUM se heurtent à un mur qui est celui de l'intelligibilité des données et des représentations : les données ne parlent pas d'elles-mêmes. Le programme que trace le WUM n'a pas vocation à refermer l'analyse des usages et des pratiques sociales sur les données. Son objectif est celui d'une mise en ordre à des fins exploratoires des différentes techniques informatiques et statistiques afin de répondre le mieux possible aux besoins d'analyses quantitatives et qualitatives qu'introduit une approche par les données. Aller au-delà de cette réponse ne peut s'envisager sans un rapprochement avec d'autres méthodes d'analyse et d'autres angles d'approche de l'usage portés par les disciplines concernées par le champ d'étude des usages du Web.

Dans cette nouvelle étape d'un programme appelant l'interdisciplinarité, il s'agit de mobiliser des connaissances de nature qualitative permettant d'interpréter des manifestations comportementales et des phénomènes observés dans les termes des usages sociaux du Web.

Du point de vue des Sciences sociales, le recours à l'expérimentation donne le sentiment d'un développement de pratiques hétéronomes, éloignées des préoccupations scientifiques et des problématiques de leurs domaines. La prise en compte de nouveaux types de données, telles que les traces d'usage, ne doit cependant pas être négligée mais, au contraire, intégrée dans un processus critique qui vise à soumettre les résultats intermédiaires, à une lecture propre à en pointer les défaillances, les limites comme les avantages.

Dans l'élaboration permanente du champ disciplinaire des Sciences de l'information et de la communication, le concept de trace parait avoir un statut délicat. Comme l'énonce Alexandre Serres, les Sciences de l'information et de la communication ne peuvent pas se confondre avec les *disciplines indiciaires* (Serres, 2002, p10). Historiquement, les SIC se sont constituées dans le courant d'une pensée en Sciences humaines et sociales du 20^{e} siècle fortement théorique, situant ses enjeux dans une compréhension globale de la société. Dans ce contexte, les lectures les plus légitimes de

la *trace* sont associées d'une part, à l'explosion médiatique et son effet panoptique dans l'évolution des formes de la communication et, d'autre part, à la traçabilité des individus comme fait de société (Mattelart, Vitalis, 2014). Il s'agit d'une définition de la trace en tant que symptôme d'une réalité à étudier. L'approche documentaire de la trace est également légitime d'un point de vue méthodologique, comme moyen d'investigation local, à un échelon *micro*, pour éclairer des réalités ponctuelles.

La critique principale adressée par les SIC et les SHS dans leur ensemble au concept de trace d'usage est relative à la question de la généralisation et de l'abstraction conduite à partir de collections indicielles. Le fondement méthodologique de cette critique est tout à fait légitime mais cet argument massif traduit parfois également une volonté de démarcation vis-à-vis de pratiques perçues comme extra-disciplinaires et n'ayant pas vocation à restituer les réalités humaines et sociales.

Une évolution se dessine cependant qui tend à habiliter progressivement la notion de trace numérique d'usage dans les recherches en SIC. Deux axes se dégagent. Le premier renvoie aux problématiques de l'identité dans les médiatisations et les processus info-communicationnels à l'œuvre sur le Web (Merzeau, 2009, 2013). Le second est de nature plutôt sémiologique et tend à intégrer les traces numériques dans une réflexion globale sur le signe (Galinon-Mélénec, 2011), (Galinon-Mélénec, Zlitni, 2013). Dans ces différents travaux, la trace numérique est un objet autour duquel se nouent des enjeux sociaux (réappropriation, etc.) mais elle n'est pas convoquée en tant que donnée de traitements.

Nous reviendrons sur ces différentes questions de nature à la fois méthodologique et épistémologique dans la seconde partie de ce mémoire.

Partie II

Préambule à la deuxième partie

Produire des données est un mot d'ordre autant qu'un enjeu parfaitement admis au sein des Sciences humaines et sociales. Cette activité est considérée indispensable dans l'élaboration d'un projet scientifique. Elle suppose l'existence d'un objectif duquel découle un programme organisé suivant le choix d'une méthode en différentes étapes conduisant l'élaboration progressive de connaissances associées à l'objet de recherche. Selon le modèle scientifique empirique, produire des données est l'une de ces étapes. Celle-ci est précédée d'une séquence organisant la production et elle est suivie d'une séquence organisant leur exploitation. Ainsi, les données se trouvent être un point d'articulation entre deux séquences coordonnées. Dans le schéma scientifique classique, cette coordination est maîtrisée ce qui signifie que la production est totalement guidée par la définition préalable de la séquence suivante. Les données entrent en nature et en volume dans un cadre fixé par les outils et les méthodes d'analyse. Celles-ci régissent également les conditions de leur production. Cette dépendance forte des données vis-à-vis des conditions de leur exploitation laisse peu de place à l'imprévu, si ce n'est dans la difficulté (voire l'impossibilité) de leur production ou dans l'interprétation des résultats d'analyse.

Cependant, dans le cas des données numériques, le schéma classique que nous venons de présenter ne répond cependant pas à toutes les situations de production de données qui sont/seront exploitées. L'exploration de situations nouvelles ou de dispositifs complexes, ne permet pas toujours d'avoir une idée exacte et précise de la nature des données ni de quelles manières celles-ci pourront être exploitées. C'est précisément cette situation qui nous a conduit à nous intéresser à la réalisation d'un objet représentationnel intermédiaire partiellement finalisé et que nous désignons par *collection*. Il s'agit d'une structure de données agrégative, organisée suivant une dimension temporelle et contenant potentiellement des données analytiques. L'objectif associé aux collections est de maintenir une cohérence méthodologique dans des situations ou les processus de production et d'exploitation n'ont plus d'articulation évidente.

La numérisation de l'environnement dans lequel nous vivons nous a familiarisé avec les objets informatiques que sont les *données structurées* qui soutiennent les traitements et les services info-communicationnels. Les pages Web, les cookies, etc. qui circulent ou s'enregistrent dans nos outils connectés, contribuent à naturaliser les données numériques. Pourtant, produire des données fait bien référence à un processus d'élaboration.

Contrairement à ce que suggère le terme, il ne s'agit pas de récolter des données existant spontanément et uniformément dans l'environnement mais bien de les fabriquer.

L'immédiateté de l'information numérique à laquelle nous accédons en quelques clics fait illusion. En particulier, elle peut laisser croire que nous disposons de ce que nous voyons, ce qui n'est pas le cas. La page vue dans le navigateur n'est que la projection graphique d'un résultat de traitements d'autant plus complexes que les fournisseurs qui opèrent le service d'information ne veulent pas que l'on puisse capturer les données qui la constituent.

Symétriquement, l'activité numérique à laquelle nous n'échappons plus dans l'organisation de notre quotidien alimente le fantasme d'une omnipotence du numérique qui traduirait dans ses moindres détails ce que nous faisons, ce qui n'est pas non plus le cas. Le traçage numérique relève de modèles fragmentaires dans des espaces fortement concurrentiels ne favorisant pas l'unification des représentations.

Ainsi, produire des données implique de s'engager fermement dans une ingénierie informationnelle favorisant autant que possible, dans les solutions retenues, ce qui est généralisable sans pouvoir cependant échapper aux singularités. Toute observation va alors comporter un processus technique d'adaptation et de mise au point de la production de données intégrant les contraintes marchandes et juridiques d'un espace numérique qui, s'il est largement ouvert au public, se privatise dans l'accès aux données de services. Il s'agit alors de composer avec les filtres apposés par les opérateurs qui régulent l'accès pour leur compte ou au nom d'intérêts supérieurs. De très grand déséquilibres se font jour entre les différentes communautés scientifiques internationales. À titre d'exemple, la récupération de données de Twitter auprès de la société GNIP, pour un hashtag et une période de 3 ans (2010-2013) a été évalué à 15 750 dollars pour environ 400 000 Tweets[137]. Dans le même temps, la décision prise en 2010 par Twitter de reverser au travers de la société GNIP à la bibliothèque du congrès l'intégralité de ses données[138] (plusieurs centaines de milliards de Tweets) traduit bien le malaise auquel nous sommes confrontés.

Ce constat soulève la question fondamentale des limites de ce qu'il est possible techniquement et déontologiquement de faire dans l'espace numérique. Mais au-delà, c'est la définition politique de l'existence d'un espace d'accès ouvert (public) aux données qui se pose.

Définir des collections de données est peut être aussi une manière d'apporter une réponse aux enjeux économiques et politiques de la production des données. Cela ne devient possible que si les logiques de production de collections s'inscrivent dans les normes et les processus documentaires de l'*open Data.*

[137] Devis obtenu en préparation d'un projet préparé en 2017 contenant une phase d'analyse rétrospective de Twitter (ANR *HOSPICITE*. cf. chapitre 6).

[138] Depuis 2006.

CHAPITRE 4
Approches expérimentales et dispositifs de traçage

« *What happens if the subject see the experimenter" (behind the "curtain" in an adjacent room acting as the computer)*?
Kelley answered, "*Well, that's just like what happened to Dorothy in the Wizard of Oz.* »
And so the name stuck.
(Green, Wei-Haas, 1985)

Dans ce chapitre nous revenons sur les trois *dispositifs* d'observation expérimentale que nous avons conçus et réalisés, et qui ont été mis en œuvre dans le cadre de contrats de recherches interdisciplinaires. Nous utilisons ici le terme dispositif pour rendre compte de la cohérence des configurations logicielles que nous installons (sondes, etc.) ainsi que des canaux d'information qui nous organisons afin de constituer le traçage du dispositif observé.

Les dispositifs présentés ont été réalisés à une dizaine d'années les uns des autres. SOPHOCLE, le premier répondait encore à une logique de poste de travail alors que les deux suivants se sont inscrits dans l'évolution des technologies info-communicationnelles du Web (PLEXUS) puis des réseaux sociaux (MEDIASWELL). Pour autant, il ne s'agit pas d'une déclinaison de réponses successives motivées par l'évolution technologique continue dans le cadre d'un programme fixe. Les programmes scientifiques qui sous-tendent chacune des réalisations n'ont pas la même visée. Ils intègrent chacun un questionnement qui est celui d'une époque et d'une discipline elle-même confrontée à l'évolution socio-technique de ses pratiques.

À l'issue d'une présentation séparée, nous proposerons une synthèse dont l'objectif est d'articuler les éléments méthodologiques qui fédèrent ces trois réalisations.

Afin d'unifier les présentations, nous utiliserons deux types de graphiques.

Le premier que nous qualifions de *Schéma de configurations*, décrit au travers d'une typologie (sujet, dispositif, ressource) les distributions d'acteurs et les faisceaux d'interactions qui font l'objet du programme d'observation associé au dispositif expérimental. Dans cette typologie, les différents opérateurs de l'expérimentation sont confondus avec le dispositif qui les masque. De manière symétrique, les différents objets interactifs réalisant l'instrumentation interactive du sujet sont confondus avec le sujet qui les utilise. On entend par *ressource* toute externalisation informationnelle (base de données, logiciel, système d'information, etc.) dont l'utilisation est requise à l'accomplissement de l'activité du sujet ou de l'expérimentateur. Ces schémas généraux ont vocation à s'instancier suivant un protocole de réalisation expérimental. L'étendue du dispositif d'observation mis en œuvre est matérialisée par les fonds blancs des objets du schéma, les fonds

grisés correspondant à des parties non accessibles. L'étendue complète d'un acteur traduit une intention qu'il convient de distinguer de sa réalisation qui ne peut être qu'approchante.

Le second décrit l'*architecture fonctionnelle* du dispositif d'observation. Les boîtes correspondent à des unités fonctionnelles ou modules qui regroupent un ensemble de fonctionnalités opérant sur des objets communs ou contribuant à une même étape de traitement. Les doubles flèches entre les modules traduisent les dépendances structurelles associées au fonctionnement du dispositif.

1. Simuler des comportements : SOPHOCLE

1.1. Le paradigme du *Magicien d'Oz*

La désignation de *Magicien d'Oz* remonte au début des années 1980. Elle est attribuée à Jeff Kelley (alors doctorant) à l'occasion d'un séminaire du laboratoire de recherche sur les communications du département de psychologie de l'université Johns-Hopkins (Baltimore) dirigé par Alphonse Chapanis. Alors qu'on lui posait la question de savoir ce qui se passerait si la supercherie expérimentale était révélée, il utilisa la métaphore de la découverte du charlatan par Dorothy dans l'œuvre de Lyman Frank Baum (1900) (Green, Wei-Haas, 1985, p1). L'origine des travaux que cette référence étiquettera désormais, est en fait plus ancienne (milieu des années 1970) mais correspond néanmoins aux travaux de l'équipe dirigée par A. Chapanis.

Les motivations de ce type d'expérimentations toujours actuelles sont à la fois économiques et technologiques. En effet, la simulation permet de tester des hypothèses innovantes et des scénarios d'usage dans un délai et un coût raisonnable et dans des cadres techniques allant parfois au-delà des capacités contemporaines des technologies. Le paradigme associé est présenté comme celui d'un expérimentateur intervenant dans la boucle technique d'un dispositif interactif simulé. Le principe est qu'un ou plusieurs expérimentateurs que nous désignons comme *compère(s)* effectue(nt) un sous ensemble de tâches dans un processus supposé être accompli intégralement par la machine à laquelle le sujet est confronté. Dans une définition *forte* du paradigme[139], la simulation ainsi que le contexte expérimental ne sont pas connus de l'utilisateur et sont masqués afin de maintenir l'illusion d'un fonctionnement vraisemblable d'une machine autonome.

La *vraisemblance* est un point clef des dispositifs expérimentaux de ce type qui sont le plus souvent envisagés dans un contexte d'innovation très marqué, voire de rupture. En premier lieu, elle consiste à gommer les imperfections que l'intervention du compère introduit dans le processus de traitement. Celles-ci sont le plus souvent liées à la capacité de traitement et à la durée d'exécution du processeur humain. Pour masquer ces faiblesses susceptibles de trahir la simulation, des raccourcis sont mis en place et un système de justification est produit[140]. En second lieu, elle consiste à entretenir l'illusion en naturalisant les fonctionnalités ou les processus innovants au

[139] Qu'il est possible d'affaiblir suivant les objectifs expérimentaux.

[140] Dans le cas de SOPHOCLE, la vraisemblance a été soutenue par des boîtes de dialogue (pop-up) signalant les différentes phases et le degré de progression du processus en cours. Des messages préenregistrés permettaient également de normaliser les réponses.

travers de rétroactions (*feedback*) renforçant les représentations mécanistes du sujet. Par exemple, l'entraînement préalable d'un module vocal simulé, comme l'affichage mot à mot des transcriptions dactylographiées de séquences produites, assure l'utilisateur de l'existence d'un module de reconnaissance vocale tout en lui donnant le contrôle sur l'intelligibilité de ce qu'il a produit. Cette naturalisation de la simulation est délicate. Elle peut même en constituer un biais dans la mesure où elle repose sur des hypothèses de continuité avec le contexte technologique contemporain.

Au travers du Magicien d'Oz nous mettons en évidence l'une des clefs de notre approche. Il s'agit de la mise en œuvre d'un principe écologique, c'est-à-dire d'inscrire le processus d'observation dans l'environnement et l'écosystème des sujets.

Un autre élément nous parait important à souligner. Celui-ci est historique et porte sur le lien fort qu'entretiennent, dès cette époque, nos études avec le paradigme de performance qui est ici sous-jacent à l'activité d'observation. Dans le cas présent, ce paradigme est associé à *l'usabilité* des systèmes autrement dit à leur prédisposition à l'usage. En réalisant un dispositif d'observation tracé, nous restions alors indépendant des logiques industrielles de production qui étaient à l'œuvre mais qui ne nous atteignaient pas.

1.2. Le dispositif SOPHOCLE

SOPHOCLE[141] a répondu à des attentes exprimées au début des années 1990 par les chercheurs investis dans l'étude du dialogue homme-machine et des médiations informatisées. Comme de nombreux autres travaux de l'époque, l'objectif concernait prioritairement la modalité langagière et l'écrit en particulier (Dahlbäck et. Al. 1993). Si l'étude de la modalité langagière écrite naturelle c'est-à-dire non contrainte a été une facette principale du projet SOPHOCLE, le dispositif expérimental répondait plus généralement à l'étude de toutes formes d'expression et communication écrites contrôlées par un protocole.

Ce dispositif expérimental a permis d'étudier des médiations en recréant différentes situations d'interaction caractérisées par le schéma suivant (Fig19) :

- Un sujet (S), utilisateur à son insu (le plus souvent) d'un dispositif dont tout ou partie des modalités d'interaction ou des fonctionnalités applicatives sont simulées ;
- Un dispositif de simulation (D) mis en œuvre par un ensemble de compères ;
- Un ensemble de ressources (R) synthétisées par un ensemble de compères.

[141] SOPHOCLE est l'acronyme de Système pour l'Observation des Pratiques Homme-Ordinateur Conduites en Langue naturelle Écrite. Nous renvoyons à la lecture de la Ref.2. Pour la présentation historique du dispositif ainsi qu'à l'introduction de ce mémoire pour son inscription historique dans mes travaux.

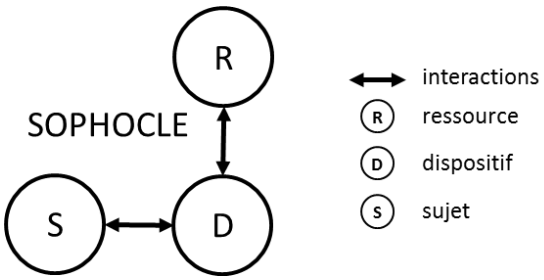

***Fig19. Schéma de* configurations de SOPHOCLE**

Suivant les déclinaisons de protocoles formalisant la conduite des interactions de (D) vers (S) et des logiques d'utilisation des ressources (D<->R), les situations expérimentales caractéristiques de productions dialogiques envisagées ont porté sur :

- La présence de deux interlocuteurs identifiés comme énonciateurs et protagonistes d'une session interactive qui leur est nommément associée[142] ;
- Des modalités d'énonciation (production d'énoncés) et d'interaction (enchaînement des tours de parole) totalement dégagées des contingences d'une application particulière.

Pour atteindre cet objectif, le dispositif se décompose en sous-parties distinctes (cf. Fig20) :

- Une partie graphique caractérisant une interface applicative à laquelle accède le sujet et lui permettant de conduire une activité ciblée ;
- Une partie communication qui permet de contrôler à distance (asservir) l'environnement graphique du sujet et de canaliser les échanges asynchrones au cours d'une *session* ;
- Une partie applicative dédiée au magicien lui permettant : d'interagir avec des ressources applicatives afin d'effectuer les actions nécessaires (*i.e.* recherche d'information) et de simuler dans les meilleures conditions des comportements d'interfaces ou de suppléer des fonctionnalités applicatives inexistantes ;
- Une partie interface avec des ressources externes mobilisables par le compère dans l'élaboration de sa réponse (ex : un serveur vidéotex) ;
- Une partie archivage dont la fonction est l'organisation et l'enregistrement des traces expérimentales des sessions afin de constituer une collection ;
- Enfin, le cœur de synchronisation événementielle, réalisant le *monitoring* expérimental et la communication entre les différentes parties (*i.e.* modules).

[142] Outre le sujet (observé) la deuxième individualité est virtuelle et assumée par les compères au travers du dispositif.

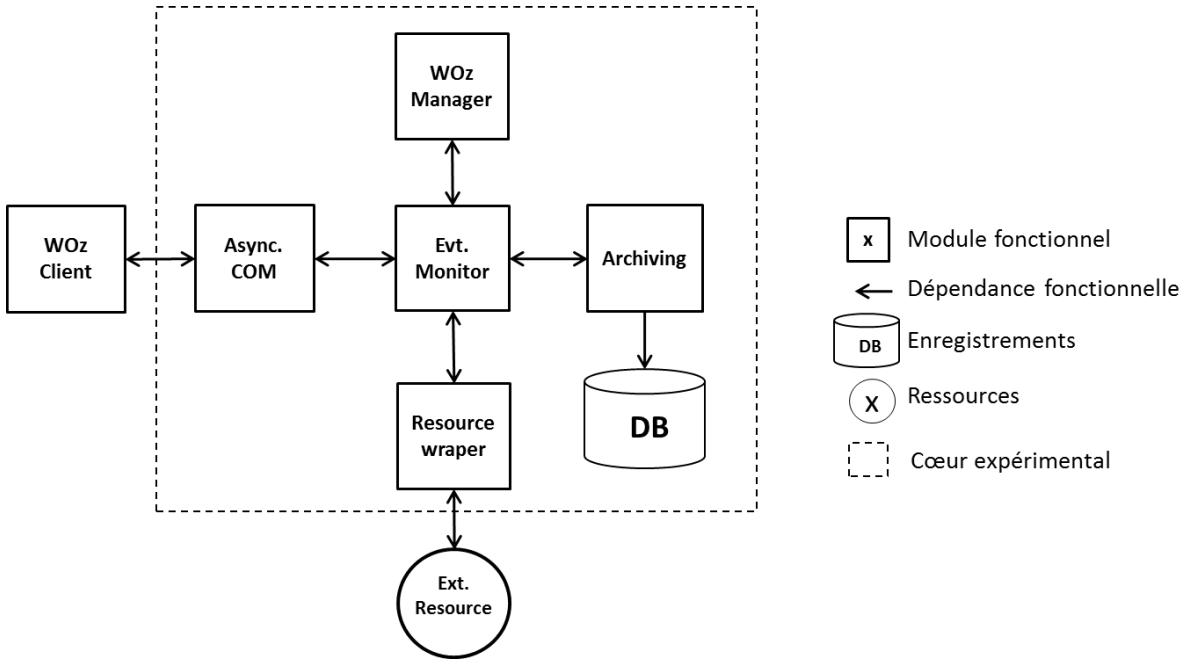

Fig20. **Architecture fonctionnelle de SOPHOCLE**

Ce dispositif expérimental a été conçu à une période où le développement des réseaux en entreprise était orienté vers l'usage interne et guidé par les solutions propriétaires proposées par les éditeurs logiciels. L'analyse à distance des interactions effectuées par un utilisateur sur un poste de travail ne pouvait pas être envisagée de manière générale en s'appuyant sur des environnements (i.e: navigateur Web) et des technologies fortement normalisées (i.e HTML, HTTP). Les développements réalisés, appuyés sur les technologies propriétaires Apple relevaient d'un choix justifié par l'ergonomie des appareils (Macintosh), autonomes et aisément transportables. Le kit SOPHOCLE se devait d'être facile à déployer en tout lieu quel que soit son degré d'informatisation. La contrainte principale était de pouvoir isoler physiquement les *compères*[143] des sujets observés (Ref.2).

Dans le contexte MMI2[144], qui lui est contemporain, des expérimentations très similaires sur le plan du protocole (simulation) ou des objectifs (collecte de dialogues) mais portant sur l'interaction multi modale libre ont été réalisées en laboratoire. Ces expérimentations utilisaient l'interface graphique mise en œuvre dans le démonstrateur du projet et un système de communication entre deux terminaux d'un même réseau. Ces fonctionnalités permettaient de répliquer à distance le poste du sujet, d'interagir en retour avec son environnement visuel et d'enregistrer les échanges.

1.3. Visées expérimentales

Les deux types d'expériences, mono ou multimodales, ont partagé des objectifs communs. Dans les deux types de situations, le but principal a été de constituer des enregistrements de dialogues

[143] Terme utilisé pour désigner les protagonistes, expert(s) métier et expérimentateur(s).

[144] Projet Européen Esprit : *Multi Modal Interface for Man Machine Interaction*. cf. chapitre 1.

comportant des expressions spontanées du sujet engagé dans une interaction libre avec un dispositif simulant une expertise (documentaire pour SOPHOCLE ou conception de réseaux pour MMI2). Les collections ainsi réalisées devaient répondre à trois niveaux de problématique :

- À un premier niveau, l'observation se rapporte au sujet, utilisateur du dispositif expert. On vise alors ses pratiques modales ou multi modales. L'objectif est au travers de recueils et d'exemples, de mieux formaliser et paramétrer les compétences spécifiques à chacune des modalités dans leur fonctionnement interprétatif. Pour la modalité langagière, la formalisation porte prioritairement sur les niveaux lexicaux et syntaxiques.
- À un deuxième niveau, ce sont les médiations expertes et le compère qui les effectue qui sont objets d'observation. L'éloignement du sujet exprimant ces attentes et la contrainte d'un canal écrit (voire multimodal) sont supposés le forcer à *éliciter* le processus métier qu'il réalise. Dans le contexte MMI2, il s'agissait d'un expert en conception de réseaux informatiques ; dans le contexte SOPHOCLE il s'agissait de professionnels de la documentation. L'objectif est celui de la formalisation des connaissances. Selon les hypothèses en vigueur, l'expertise métier est susceptible d'être verbalisée par le compère au travers des commentaires sur son action et sur ses interventions dialogales. Il doit ainsi être possible d'organiser un système de connaissances propre au domaine d'expertise associé à l'activité. Dans les expérimentations conduites, l'extraction des connaissances était plutôt abordée en référence aux modèles structurés en tâches de l'activité instrumentée (médiée). La formalisation de nature sémantique de représentations liées au domaine d'activité était envisagée dans le contexte des systèmes experts et de l'ingénierie des connaissances telle qu'elle se développait à l'époque.
- Enfin, un troisième niveau, se rapportant aux interactions dialogales, justifie à lui seul la captation des échanges dans le contexte de médiations expertes. S'il repose sur des productions considérées suivant l'un des deux niveaux précédents, celui-ci est le plus souvent envisagé isolément des deux autres. C'est notamment dans cette perspective que les phénomènes dialogiques régissant les tours de paroles ont été abordés. Dans les expérimentations menées, la conduite dialogique n'a jamais véritablement fait l'objet d'un protocole expérimental et a largement reposé sur la naturalité des conduites du compère.

Dans les documents faisant référence au projet expérimental de SOPHOCLE (Ref.1-Ref.4) les trois questions centrales sont celles : de la place (voire de sa légitimité) de la langue naturelle dans les interfaces (ILN) ; de son degré de spécialisation dans un contexte de médiation technique ; des spécifications linguistiques propres à la mise en œuvre de modules spécialisés à son traitement sont centrales. Cette centralité fait écho à des problématiques dont nous avons rappelé certains des ressorts (cf. Ch1). L'approche de l'interaction homme-machine qui en découle s'inscrit dans une logique de reformulation que l'on trouve exprimée dans le paragraphe 3.2 p5 : *le passage de la requête en LN à la requête en langage de commande*. Le paradigme exprimé est celui de la *traduction*. La machine idéale a en conséquence, pour objet de ramener l'interaction dans un champ d'actions fermé et finalisé. La compétence dialogique attendue est celle qui permettra d'opérer ce glissement de manière continue et le plus naturellement possible. Une telle conception de l'interaction fixe les enjeux expérimentaux et oriente, volontairement ou non, la mise en œuvre expérimentale. Cette

remarque souligne notamment l'importance du protocole et de l'encadrement des comportements des compères. Les consignes définies en amont et qu'il faut ensuite tenir dans le fil de l'expérimentation ont une fonction méthodologique de première importance. L'absence de protocole est en fait un protocole par défaut dans lequel interfèrent différents enjeux, notamment ceux qui ont amené l'expérimentation. Dans le cas présent, la définition est d'autant plus exigeante qu'il s'agit de compétences naturelles : la langue et le dialogue interpersonnel.
Les expérimentations de type Magicien d'Oz réalisées avec SOPHOCLE n'ont, de ce point de vue, pas dépassé la magie de l'expérimentation. Nous avons certes recueilli des dialogues, mais la diversité des formes rencontrées et les fragilités du protocole auraient dû nous conduire à raisonner l'expérience pour recadrer et spécialiser les suivantes. La simulation de type Magicien d'Oz, est exigeante pour être payante au-delà des premiers résultats. La porte que nous avons ouverte nécessiterait un investissement durable, dans une démarche empirique.

2. Les usages et pratiques sociales du Web : PLEXUS

2.1. L'individualisme connecté

Le développement des technologies du nomadisme connecté et la généralisation concomitante du Web comme espace de promotion de services info-communicationnels nous incitaient, au début des années 2000 à définir une plateforme expérimentale pour l'étude de l'usage dont PLEXUS[145] est l'acronyme.

Ce projet a été financé par l'obtention, suite à un concours, d'un budget de l'action concertée incitative ville (ACI-Ville) mise en œuvre par le Ministère de la recherche entre 2001-2004 (Ref.14). Par la suite, le dispositif PLEXUS sera déployé sur différents sites dans un contexte de collectes permanentes d'informations sur des périodes de plusieurs mois.

Dans l'évolution de notre carrière, notre projet de recherche n'est plus guidé uniquement par l'élaboration d'interfaces et de systèmes interactifs. Il s'oriente progressivement vers les problématiques de l'usage et de pratiques info-communicationnelles. L'évaluation des dispositifs supportant les médiations devient également nécessaire. Ainsi dans le texte qui accompagne le projet de l'ACI-Ville, se côtoient deux aspirations distinctes :

- Celle d'une part, d'intégrer mes travaux « *dans un champ de recherche qui porte sur les relations et les dynamiques réciproques entre société et technologies de l'information-communication* » ;
- Et celle, d'autre part, d'avoir comme objet d'étude « *l'interaction entre l'individu et les dispositifs d'information-communication dont il dispose dans le cadre du déroulement d'activités finalisées qu'il conduit.* » (Ref.13, p351).

Le premier point est une affirmation disciplinaire mais aussi la reprise des hypothèses constitutives de l'ethnotechnologie (cf. chapitre 4, §2.1) qui fonde notre démarche personnelle. Le second, révèle

[145] PLEXUS est également un substantif décrivant un réseau de nerfs qui s'entrelacent de façon complexe en un point de l'organisme. L'utilisation métaphorique s'est imposée à nous comme caractéristique des enregistrements associés aux traces d'usage collectées par le dispositif d'observation.

l'évolution non achevée d'un changement de paradigme qui confond encore l'objectif avec ce qui se révélera par la suite être une méthode.

En effet, l'époque est celle d'une transformation du paradigme de performance qui portera moins sur l'usabilité comme adaptation du service à l'usager que sur l'efficacité du service à engendrer un acte de consommation (transformation). Dans les expérimentations que nous avons conduites alors, c'est l'usage informationnel du Web en toute généralité et non les pratiques de services spécifiques que nous avons étudié (Ref.16). Les traces numériques d'usage que nous produisons sont associées à une activité dont nous ne retenons que la partie communicationnelle.

L'objectif assigné à la plateforme est alors double :

- « *étudier l'individu en situation, et plus particulièrement les stratégies et les comportements informationnels développés dans le cadre de cette activité. L'objectif poursuivi est d'évaluer l'incidence des dispositifs proposés sur l'individu et l'activité elle-même en intégrant à la fois la dimension nomade et les rythmes dans lesquels se déroule l'activité.* »
- « *proposer des modes d'analyse et de validation des dispositifs d'information-communication qui, partant d'une démarche expérimentale d'observation des utilisateurs, conduisent à des modèles prédictifs de conception et d'évaluation restituant, autant que possible, la nature écologique et les "affordances" de l'utilisation en situation.* » (Ref.14, p351).

2.2. Le dispositif PLEXUS

Le dispositif expérimental PLEXUS a été structuré afin de pouvoir équiper des outils de consultation mobiles (smartphones) ou fixes (ordinateurs) et de suivre de manière continue et asynchrone une population d'individus ainsi équipés.

Les modalités de l'observation expérimentales se caractérisent par le schéma suivant (Fig21) :

- Un ensemble de sujets (S), utilisateurs d'un navigateur et parcourant le Web ;
- Un dispositif de traçage (D), répertoriant les traces d'usage captées par des sondes monitorant[146] les navigateurs ;
- Les ressources (R) du Web qui sont associées aux navigations des internautes.

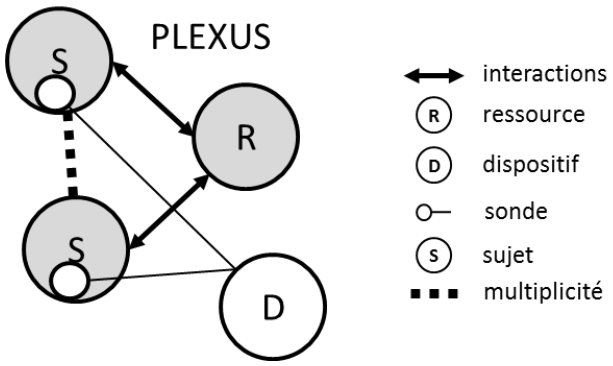

***Fig21. Schéma de* configurations de PLEXUS**

[146] En l'occurrence, les sondes réalisent un enveloppement (encapsulation) de l'exécution des clients ainsi monitorés.

Pour atteindre cet objectif, le dispositif se décompose en sous-parties distinctes (cf. Fig22) :

- Une partie *client* qui correspond à la déclinaison d'une sonde (suivant le type d'appareil) qui capte les différents événements qui surviennent dans le navigateur lors d'une consultation. Ces événements sont liés, en premier lieu, aux actions décelables de l'utilisateur sur le navigateur lui-même (déplacement dans la page avec les ascenseurs, actions sur le texte de sélection-copie, etc.) et en second lieu, au flux http que la navigation sur le Web engendre. Ce flux contient l'ensemble des éléments graphiques et ressources constituant les vues[147] (dont les bandeaux publicitaires et autres pop-up), ainsi que les documents téléchargés ;
- Une partie *serveur* qui centralise et régule la communication permanente avec les différents clients (sondes) équipant les sujets de l'observation ;
- Une partie événementielle qui analyse les différents événements et traces recensés par le serveur afin d'en proposer une représentation enrichie propre à être archivée ;
- Une partie archivage qui enregistre dans une base de données les traces complétées.

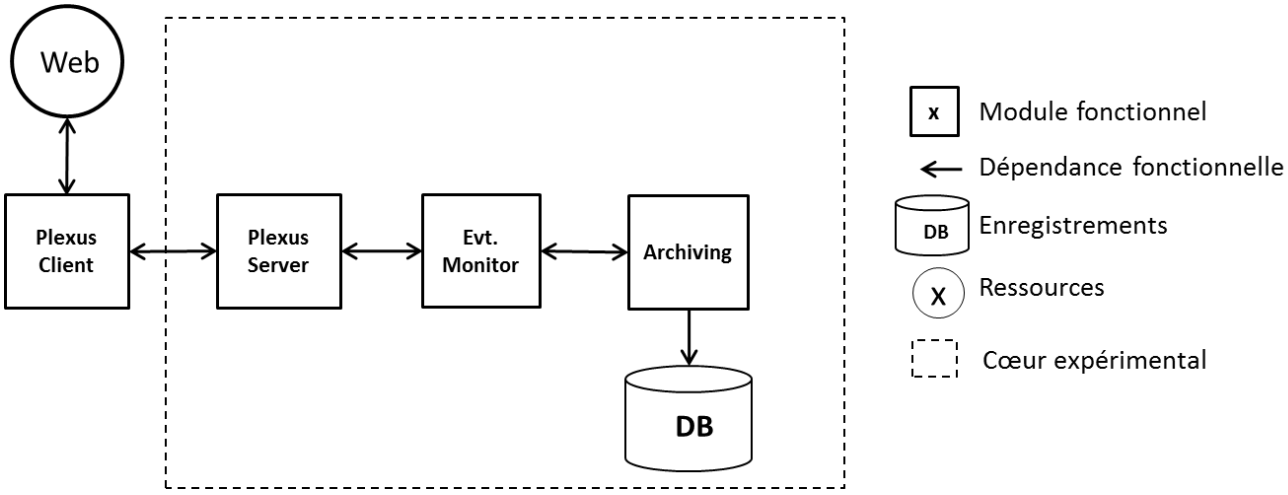

Fig22. **Architecture fonctionnelle de PLEXUS**

2.3. Visées expérimentales

L'ambition du dispositif est de permettre une *observation écologique*, c'est-à-dire in situ, respectant et restituant fidèlement, le plus finement possibles, sans altération décelable (ralentissement, etc.), de la conduite instrumentée (instrumentalisation ou instrumentation) de l'activité. Cette naturalité de l'observation établit une continuité entre les projets SOPHOCLE et PLEXUS. L'évolution réside dans le fait que la simulation n'est plus un objectif et que le compère est externe au dispositif technique et n'a plus prise ni sur la médiation, ni dans l'interaction avec le sujet.

[147] Ce flux est complexe à analyser, notamment du fait de la technologie Ajax qui associe de nombreuses ressources HTML, CSS, Javascript, etc. pour calculer à la volée une mise en forme finale de la page vue.

Les expériences réalisées avec ce dispositif se rapportent aux problématiques d'inscription des usages et des pratiques sociales dans les logiques d'organisation de l'activité humaine et des rythmes de vie.

Le protocole expérimental est de ce point de vue, plus léger. Cependant, il nécessite l'obtention du consentement de l'usager pour rendre possible l'exploitation des données de l'observation. Nous avons opté pour le principe d'une information (boîte de dialogue) au moment du lancement du navigateur qui prévient que la navigation fait l'objet d'un enregistrement qui peut être exploité dans le cadre d'une étude des usages du centre[148]. L'observation est envisagée dans la durée (plusieurs mois) sans contrôle permanent de la situation d'utilisation. Un dispositif d'analyses complémentaires (entretiens et questionnaires) est déployé ponctuellement. Les deux expérimentations majeures (2006) (Ref.23) se sont déroulées dans des espaces d'accès publics à l'Internet dans des territoires de montagne[149]. Soit une vingtaine de points de consultations fixes (pas de wifi) accessibles dans des plages d'ouverture quotidiennes. L'intérêt majeur de ce type de configuration réside dans le fait que les navigateurs Web constituent la seule instrumentation autorisée en dehors des outils de bureautique élémentaire[150].

L'unité d'observation est la session c'est-à-dire l'intervalle durant lequel un sujet qui n'est pas authentifié navigue de manière continue sur le Web. Une session s'ouvre et se referme avec la fenêtre d'exécution du navigateur. Un même sujet peut être associé à plusieurs sessions durant sa visite. Les erreurs de manipulations telles que les fermetures/ouvertures intempestives sont le plus souvent rattrapées à partir d'une fenêtre de recouvrement entre sessions[151]. Au-delà de cet intervalle de sécurité, les sessions sont considérées comme distinctes et il n'est plus possible d'attribuer les sessions à un même sujet. Le sujet est par hypothèse réduit à un seul individu.

L'analyse des sessions répond à deux types d'objectifs :

- Dans une perspective macroscopique, étudier les logiques d'organisation et évaluer l'incidence structurante (spatio-temporelle) de l'accès public Internet dans l'activité journalière ;
- Dans une perspective microscopique, étudier les logiques d'activités info-communicationnelles et les finalités des sessions.

Une caractérisation des contenus de pages vues a été établie en fonction de la nature de ces contenus et de la nature supposée de l'intérêt de l'individu qui les consulte. Sa granularité a été calibrée pour caractériser les profils de sessions et répondre prioritairement aux études d'usage de l'espace public. La catégorisation qui en découle (Webmail, blog, moteur de recherche, tourisme, etc.) fait ressortir la faible distinction opérée à l'époque entre les ressources de natures différentes, notamment entre service et contenu. Les patterns séquentiels qui se dégagent de cette approche

[148] La manière de présenter l'enregistrement se base sur les protocoles utilisés par les hotlines. La présence de deux boutons distincts permettait de finaliser l'acceptation ou le refus et de n'activer le traçage que dans le cas positif.

[149] Espace public numérique de Valcenis Lanslebourg, Cybercentre de Roquebillière.

[150] Le dispositif Plexus a été doté de fonctionnalités permettant de suivre des opérations "hybrides" au travers d'actions de sélection et de copier/coller.

[151] Seuil fixé à 90 secondes.

centrée site (cf. chapitre 3) ont permis de formuler des hypothèses sur l'existence de logiques d'usages dans les sessions (Ref.23) permettant d'autres hypothèses sur l'intentionnalité de la visite.

Ce résultat constitue à notre sens le principal apport de ce type d'expérimentation aveugle. Sans authentification de session, il n'est pas possible de formuler des hypothèses plus fines sur les profils individuels.

3. À l'écoute du Web et des réseaux sociaux : MEDIASWELL[152]

Swell signifiant "houle" est un terme utilisé dans le milieu du surf pour décrire les conditions de vagues propices à la pratique de ce sport de glisse. L'emploi de ce terme file la métaphore nautique associée au Web et aux technologies de l'information. Il traduit l'ambition de la plateforme de pouvoir capter et restituer les dynamiques et les mouvements profonds des flux informationnels et de l'activité médiatisée ou médiatique.

L'objectif de MEDIASWELL est la réalisation de *collections* d'entités numériques semi-structurées, accessibles à partir d'Internet suivant différentes modalités et protocoles de communication. Il peut s'agir de pages Web, de résultats de requêtes auprès de bases de données, de fichiers, de flux (*bitstream*) ou de structures de données produites par des services ou des interfaces applicatives (API), etc.

La définition de ce dispositif s'inscrit moins dans la continuité des travaux antérieurs que dans une prise en compte raisonnée des enjeux sociétaux et scientifiques de la disponibilité accrue de données numériques issues des médiations info-communicationnelles. Notre réflexion méthodologique découle de la mise en œuvre de ce dispositif et de l'analyse des résultats produits.

D'une certaine manière, le dispositif MEDIASWELL rend compte d'une troisième actualisation du paradigme de performance qui désormais porte sur la valeur heuristique des processus informationnels mis en œuvre par les plateformes du Web.

3.1. Enrichir et collectionner les traces numériques d'usage

MEDIASWELL est une plateforme expérimentale conçue pour répondre aux besoins de recherches empiriques nécessitant la fouille et l'analyse de données extraites du Web ou produites par des services accessibles sur Internet. La notion de donnée est ici extensive : elle est associée à un spectre étendu de types de données allant d'une simple valeur numérique au document structuré.

MEDIASWELL a été mis en chantier à partir de 2010 afin de réaliser des collectes de *tweets* pour étudier les pratiques info-communicationnelles propres aux usages de ce média. À l'issue de nos travaux sur la campagne présidentielle de 2012, il est clairement apparu (Ref.26) que l'analyse des phénomènes informationnels et communicationnels qui se développent à l'intérieur de ces espaces sociaux nécessite, pour qu'ils soient étudiés, de questionner simultanément d'autres ressources du Web. Dans le cas de Twitter, les URL diffusées dans les tweets pointent des pages du Web dont

[152] Le chapitre complémentaire suivant celui-ci propose une description technique de l'implémentation de MEDIASWELL ainsi que des illustrations de son fonctionnement.

les contenus propres voire les sites dans leur ensemble constituent une clef d'analyse pour comprendre l'acte de publication.

Or, l'exploitation *ex post* des traces numériques se déroule dans un temps ou ces ressources risquent sinon d'avoir disparu au moins d'être dégradées (pages supprimées par exemples). En conséquence, préserver les éléments essentiels de ces ressources devient un enjeu corollaire de la collection. La question fondamentale qui est alors posée - et sur laquelle nous reviendrons dans prochain chapitre - est celle d'identifier *ex ante* ces ressources et le périmètre des éléments pertinents qu'elles contiennent.

Du point de vue de la collecte, le maintien d'un contexte interprétatif nécessite d'articuler entre eux les systèmes référentiels des différentes ressources. Or ces systèmes ne sont que très rarement interopérables. Cela signifie qu'il faut établir ces articulations à partir de connaissances externes telles que des dictionnaires, des ontologies ou des tables de correspondances qui nécessitent un investissement pré-expérimental important. Un système référentiel commun peut être envisagé à partir : d'une ligne temporelle (*timeline*) ; d'un système d'identifiant de ressources en réseau (URL) ; de normes référentielles institutionnalisées ou non (folksonomies de *hashtags* par exemple) ; de noms propres ou plus largement de désignations communes (expressions littérales). Les références communes, résolues dans les espaces qu'organisent ces ressources autonomes déterminent les éléments susceptibles de constituer un contexte interprétatif.

3.2. Le dispositif MEDIASWELL

MEDIASWELL peut être vu comme une plateforme expérimentale générique, dont la logique de conception repose sur la mise en œuvre de briques fonctionnelles autonomes et aisément paramétrables dans le but de produire rapidement des configurations de traitements adaptées aux besoins expérimentaux. Cette définition actuelle est le résultat d'une volonté de donner du sens à l'activité d'instrumentation dans nos recherches et dans la mise en cohérence des réponses techniques[153]. En référence au modèle CRISP-DM[154], le cœur de MEDIASWELL est dédié aux phases de compréhension et de préparation des données (Wirth, Hipp, 2000).

De manière simplifiée, ce dispositif expérimental a été structuré afin de pouvoir adresser différents types de ressources du Web.

Les modalités de l'observation expérimentales se caractérisent par le schéma suivant (Fig23) :

- Un ensemble de sujets (S), utilisateurs de services authentifiés sur le Web ;
- Un dispositif de traçage (D), répertoriant les traces d'usage captées par des sondes associées aux différents types de ressources monitorées ;
- Différentes sortes de ressources (R) associées au Web et aux réseaux sociaux.

[153] L'annexe associée à ce chapitre contient les justifications des choix techniques.

[154] Voir également ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/fr/CRISP-DM.pdf (consulté le 10/06/2016)

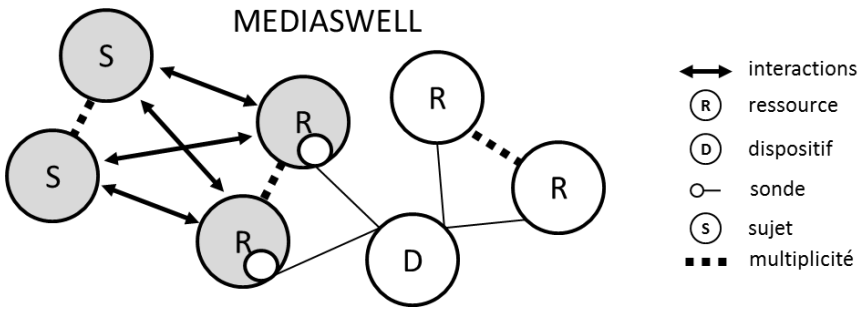

Fig23. Schéma de configurations de MEDIASWELL

Nos projets nous ont conduit à mettre en œuvre différentes compositions possibles de modules dont le fonctionnement est scénarisé. Les scénarios sont élaborés à partir de scripts d'exécution des modules[155]. Nous avons par exemple, réalisé une archive documentaire (documents PDF) associée à des publications publiques disséminées sur le Web et référencées dans une base bibliographique en ligne[156]. Des scénarios plus complexes ont été réalisés dans le contexte de collectes à l'occasion d'événements médiatiques. Nous attachons de plus en plus d'importance à la mise en œuvre de projets conduits dans une perspective que nous qualifions de *juste à temps* c'est-à-dire amenant à produire une représentation instantanée de phénomènes saisis à la volée ou résultant de l'analyse des collections en temps très limité. Ces conditions particulières sont celles de l'événement et des médias. Fournir les éléments d'une analyse au fil de l'eau est un enjeu que nous estimons très important dans le devenir de nos travaux.

La mise en œuvre de différents scénarios définit le cadre d'organisation de la plateforme. Mais les fonctionnalités qui s'imposent pour les satisfaire doivent répondre, dans leur implémentation et leur fonctionnement, aux contraintes de l'approche expérimentale décrite par trois principes généraux :

(1) *Une flexibilité et un déploiement rapide.* La conception de la plateforme doit favoriser la mise en œuvre de différentes configurations de fonctionnalités et de programmes de collectes en fonction d'objectifs expérimentaux variés (voir Annexe). La réponse apportée se traduit dans la structuration fonctionnelle adoptée (§2) et dans l'architecture hautement modulaire (§3) ;

(2) *Une lisibilité et un paramétrage fin des variables et des processus expérimentaux*. La plateforme doit favoriser la compréhension du déroulement des processus expérimentaux en cours, en restituant sous une forme la plus lisible possible, les paramètres associés à ces processus. Il s'agit ainsi d'assurer la maîtrise de l'environnement et des conditions expérimentales afin de contrôler et de suivre un protocole. L'utilisation de fichiers de configurations permet de dissocier les paramètres clefs de l'implémentation logicielle des modules. Ces fichiers font partie de la définition des conditions de fonctionnement du dispositif (notamment durant la collecte) et participent du protocole expérimental ;

---

[155] Voir le complément au chapitre 4 §2.

[156] Soutien méthodologique à la participation du laboratoire PACTE au projet ADEME, programme "*observation de la recherche sur la ville durable*" en collaboration avec les universités de Lausanne, Montréal et Stanford.

(3) *Un traçage et une documentation fine des processus et des collections ou corpus*. La plateforme doit associer un ensemble paramétré d'informations et de documents journalisés contextualisant les collections ou les corpus produits. Le principe d'une documentation est mis en œuvre dans la définition des enregistrements constitutifs des collections ou des corpus. Cette documentation des données à leurs différents niveaux de structuration correspond à ce que Jean-Michel Salaün, à la suite du collectif Roger Pédauque, définit comme *redocumentarisation* (Pédauque, 2006), (Salaün, 2007). Des clefs d'identification et de catégorisation permettent la constitution d'unités documentaires[157]. Ces unités sont enrichies des traces d'exécutions journalisées (*logs*) associés aux modules mis en œuvre durant les cycles d'activité (§2). Le *monitoring* global de la plateforme produit également des données liées au déroulement programmé des processus associés à la constitution des collections.

Ces principes reportés sur les scénarios d'utilisation nous conduisent à une structuration fonctionnelle cohérente avec les différents types d'activité engagés.

3.3. La structuration fonctionnelle de MEDIASWELL

La démarche empirique appelant d'incessants va-et-vient entre hypothèses et analyses, la structure de MEDIASWELL a été adaptée en fonction. Cette plateforme intègre une logique incrémentale dans son mode de fonctionnement. Le processus expérimental est ainsi décomposé en quatre cycles d'activités logiquement structurés. Ces cycles s'enchaînent suivant une progression régie par le traitement des données. À chaque étape de traitement ou en fonction du résultat obtenu à l'occasion de celle-ci, MEDIASWELL permet de revenir sur un état antérieur des données et de reprendre les cycles des traitements jusqu'à ce que le processus expérimental converge.

Ces cycles sont coordonnés par les données qui intègrent en fonction des éléments permettant l'historisation des étapes du traitement et le tracé de l'exécution. L'unité chronologique ainsi que la cohérence des collections et des corpus en cours d'élaboration sont assurées par des métadonnées de contrôle établies durant les traitements. Ces métadonnées autorisent l'élaboration itérative, la mise en œuvre de versions alternatives et la rétroaction.

Les quatre cycles d'activités font l'objet d'une présentation détaillée. Nous les identifions chronologiquement comme suit :

(1) Définition des collections et collecte des données sources ;
(2) Constitution de la collection de données *premières ;*
(3) Mise en œuvre d'un enrichissement ;
(4) Supervision et régulations externes.

[157] Dans ce contexte, il s'agit de regroupements de représentations structurées constituant une entité pertinente en réponse à des considérations documentaires ou informationnelles dans le contexte de l'expérimentation (collecte ou analyse).

3.3.1. Cycle 1 : définition des collections et collecte des données sources

Nous employons le terme de collection dans la continuité de son usage dans le domaine de l'archivage du Web[158]. Ce cycle consiste à définir et réaliser des collections de représentations semi-structurées. Une représentation structurée désigne toute expression (textuelle) valide produite par un langage formel (dialecte XML, JSON, CSV, etc.). Le terme semi-structuré signifie que ces représentations ne sont que partiellement (voire imparfaitement) structurées. Cela ne signifie pas que les données sont incohérentes mais que leur cohérence formelle n'est pas complètement décrite et garantie. Nous qualifions de *sources*[159] les données d'entrées de MEDIASWELL (*input*) provenant de l'Internet. Ce type de données s'étend désormais aux flux numériques délimités (*bitstream*).

L'évaluation de l'adéquation des données au projet de collecte est complexe. L'écart peut être important entre l'impression qui se dégage d'une ressource consultée (une page Web par exemple) et l'extraction d'information qui lui est associée, même lorsque celle-ci est encadrée par une interface d'accès aux données (API).

La constitution d'une collection répond au besoin d'établir une représentation associée à un état momentané d'une ressource de l'Internet. Il s'agit d'élaborer une vue instantanée et représentative, à la manière d'un cliché photographique, saisissant dans un instant particulier des données ou des documents dont on ne peut pas garantir la stabilité ultérieure. L'enregistrement isole ces données provenant d'un cycle de vie qu'on ne maîtrise pas. De cette manière les traitements effectués par la suite pourront être opérés voire reproduits sur des données contrôlées, stables et permanentes. L'enregistrement permet en outre de gagner en efficacité et en performance sur les traitements effectués. L'isolement n'est cependant pas instantané. Il demande une durée qui est celle de la réalisation de la collection. Nous sommes conduits à admettre la parfaite stabilité des données sources dont l'image (la représentation originelle) n'est capturée qu'une seule fois durant la collecte.

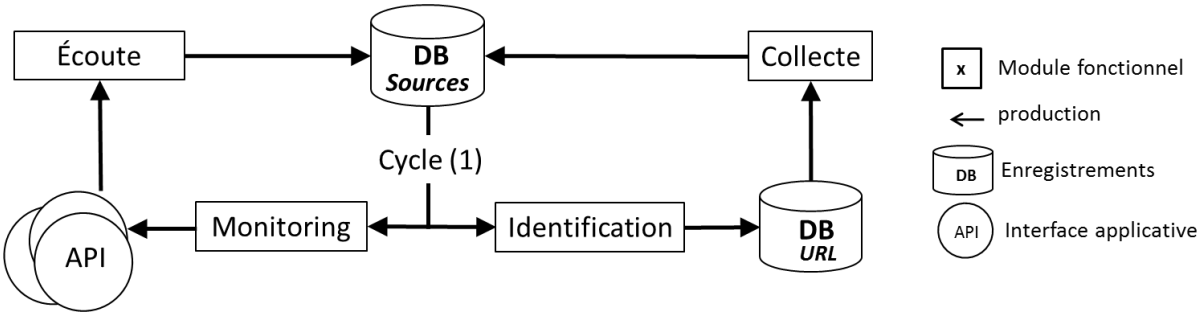

Fig24. Cycle1 - Identification des ressources

L'identification de ressources peut être externalisée ou réalisée à partir de modules proposés dans la plateforme. Dans la plupart des cas, l'identification porte sur des URL associées à des contenus publicisés. L'identification peut devenir un projet de collection en lui-même, notamment dans le

[158] Dans ce domaine, une collection est associée à un projet d'archivage systématique et non exhaustif d'un ensemble de pages du Web représentatives des logiques éditoriales associées aux ressources dont elles sont extraites.

[159] Le terme *source* est proposé comme traduction du terme anglais *sourcey* proposé comme alternative au terme brut (*raw*). Les données de source n'ont aucune garantie de cohérence, de complétude et de pertinence.

cas d'une veille préalable à l'orientation d'une collection événementielle (dépendante par exemple de l'actualité médiatique).

Dans le cas de thématiques stables ou de ressources encadrées, deux types de méthodes accompagnent le processus d'identification :

- La première repose sur la *fouille structurelle du Web* (WSM). À partir d'une liste d'URL graines (*seed*), il s'agit de parcourir les pages Web et de récolter, en fonction d'une stratégie exploratoire et de critères de sélection, des URL valides. Le mode de sélection employé est le plus souvent restreint à l'analyse littérale de l'URL (nom de domaine ou chemin)[160]. La vérification éventuelle des contenus intervient dans un second temps. Ce processus exploratoire sélectionne des URL qui sont enregistrées et organisées en collections.
- Le second concerne l'interrogation de moteurs de recherche (Bing, Google, etc.). Les moteurs de recherche peuvent être utilisés suivant deux modalités à partir d'une requête : conserver les URLs des pages de réponse ; utiliser les contenus des pages réponse comme évaluation d'une requête portant sur une URL.

L'instrumentation du processus d'identification va souvent de pair avec la réalisation d'une collection volumineuse d'URLs. C'est par exemple le cas dans la recherche que nous avons conduite sur LinkedIn où nous avons collecté plus de 100 000 profils d'individus supposés journalistes. La pertinence des URLs collectées via des services du Web (Bing, Google, etc.) n'est pas évidente. Elle s'évalue en fonction de la qualité du formalisme de requête proposé par ces services et de celle des réponses produites par ces mêmes services. Par exemple, les services de Bing assurent un volume de réponses constant au détriment de la pertinence de ces réponses. La vérification de la pertinence des contenus ciblés est engagée dès cet instant et se poursuit tout au long du processus de collecte.

Collecte et écoute des représentations sources

La collecte est réalisée à partir de différents types de modules caractérisés en fonction des modalités d'accès à la ressource. On distingue de cette manière, les modules mettant en œuvre un protocole de l'Internet et du Web (http, ftp, etc.), des modules dialoguant avec une API.

Dans le cas des réseaux sociaux (LinkedIn, Twitter, etc.) il est fréquent que les deux modalités (protocole http et API) puissent être envisagées simultanément. Le choix de l'une ou de l'autre dépend de la nature de l'information que l'on souhaite récolter et des restrictions opérées dans les API. En effet, la majorité des API proposées pour les réseaux sociaux sont dites *ego centrique*, ce qui signifie que seuls les individus connectés ont accès à leurs propres données. L'accès via le protocole HTTP ou *scraping*[161] devient donc nécessaire. Il permet de contourner cette restriction en exploitant les pages publiques des profils de ces individus[162].

---

[160] Les travaux que nous avons conduits à partir des métadonnées d'archivage produites par la Bibliothèque nationale de France, soulignent l'intérêt de l'analyse textuelle associée aux chemins (path) en particulier pour les sites privilégiant les billets informationnels (blog, etc.).

[161] Signifie littéralement "gratter".

[162] La publicisation des profils est du ressort des individus, l'information ainsi publicisée n'échappe pas pour autant aux règles de la propriété intellectuelle.

Pour mémoire, une interface applicative (API) est un accès, organisé par un prestataire de services, à une sélection de ses données. Dans le cas d'une collecte via API, un processus de *monitoring* doit être mis en place. Celui-ci se définit comme l'ensemble des traitements à mettre en place pour calibrer et réguler les flux de données aux bornes de l'API pour optimiser la collecte tout en respectant les modalités imposées par le service. Le *monitoring* intervient sur des variables de configuration du dispositif et plus particulièrement des sondes. Nous parlons d'écoute pour caractériser l'effet de ce paramétrage sur la réception du flux de traces d'usage. Chacune des actions de *monitoring* fait l'objet d'une description qui est associée au flux et qui participe du processus documentaire de l'expérimentation[163].

3.3.2. Cycle 2 : Constitution de la collection de données premières

Il s'agit de constituer des collections de données structurées enrichies à partir des enregistrements appartenant à des collections brutes de données sources. On qualifie ces données calculées de premières. Elles peuvent n'être que le résultat d'une traduction d'une donnée source dans un format adapté, ou être le résultat de traitements complexes à partir de données sources ou enfin la simple vérification des données capturées.

Avec ce second cycle de traitements, on aborde le processus de fouille de données tel qu'on peut le définir à partir du modèle CRISP-DM (*Cross Industry Standard Process for Data Mining*).

Ce cycle est dominé par les objectifs de l'analyse engagée. Ceux-ci sont établis dans le cadre d'un projet expérimental qui ne permet pas, le plus souvent, d'avoir une approche très précise des données attendues, ni en qualité ni en quantité. Cette incertitude est inhérente à la démarche exploratoire qui continue de s'élaborer durant la constitution des collections. Les objectifs expérimentaux s'affinent alors par la compréhension des données sources. Compte tenu de la taille ou de la complexité possible de la structure de ces données, elles peuvent être illisibles. Il faut donc introduire les moyens d'une intelligibilité de ces données. Celle-ci s'obtient à partir d'indicateurs qui permettent de vérifier que la donnée est intègre et correspond à ce que l'on attend. Par exemple, l'archivage des pages Web est un processus aveugle. On ne peut en évaluer le résultat qu'après une extraction de contenus ou la visualisation des pages archivées.

D'un point de vue fonctionnel, ce cycle regroupe l'ensemble des traitements spécialisés (modules) qui s'exécutent de manière asynchrone à partir des données sources pour constituer ou enrichir des données premières.

L'activité des différents modules du second cycle contribue à positionner différents indicateurs permettant d'évaluer la qualité de la collection de données premières. L'analyse de ces indicateurs permet de reprendre ou non certaines étapes du premier cycle, voire d'identifier d'autres ressources si nécessaires. Cela constitue en soi un processus d'amélioration continue garant de la qualité des collections.

---

[163] Ce point nous est apparu comme particulièrement important. Rendre compte d'un état de fonctionnement est une caractéristique que nous associons à la fonction expérimentale du dispositif.

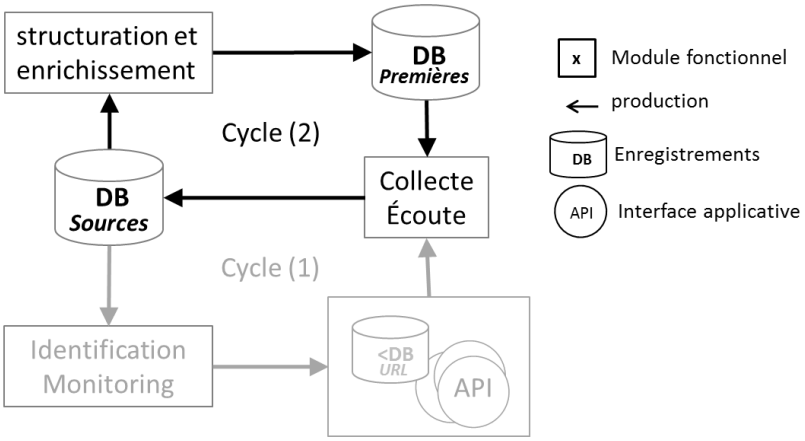

Fig25. Cycle 2[164]

3.3.3. Cycle 3 : Mise en œuvre d'un enrichissement

Ce cycle correspond à la production de données analytiques et de corpus de données susceptibles d'être exportés dans les différents formats et standards d'outils d'analyse ou de visualisation. Son objectif est de constituer la matière pour des analyses et des interprétations abstraites. Les processus analytiques viennent enrichir les représentations premières ou produire de nouvelles données, considérées comme *secondes*.

Nous désignons par enrichissement le cycle des traitements qui, s'appuyant sur les données disponibles, calculent des indicateurs, définissent des catégories et réalisent des regroupements (*clusters*) d'enregistrements de données premières.

Les traitements mobilisés à ce niveau dépendent du projet d'analyse et des collections. Ils ne permettent pas de concevoir des modules mais une boite à outils de fonctionnalités associées à des techniques de la fouille de données (machine learning, classification, clustering, etc.).

Il en va de même pour les fonctionnalités d'export qui dépendent de la structuration des données exportées (graphes, etc.) et des formats de leur enregistrement. L'exportation est considérée également comme une étape du traitement durant laquelle les données sont envisagées en tant que corpus. C'est pour cette raison que le cycle 3 se referme sur les données premières qui sont estampillées afin de les dater et d'identifier les corpus auxquels elles appartiennent.

[164] Les deux facettes du cycle 1 sont repliées sur elles-mêmes pour faciliter la lecture des cycles suivants.

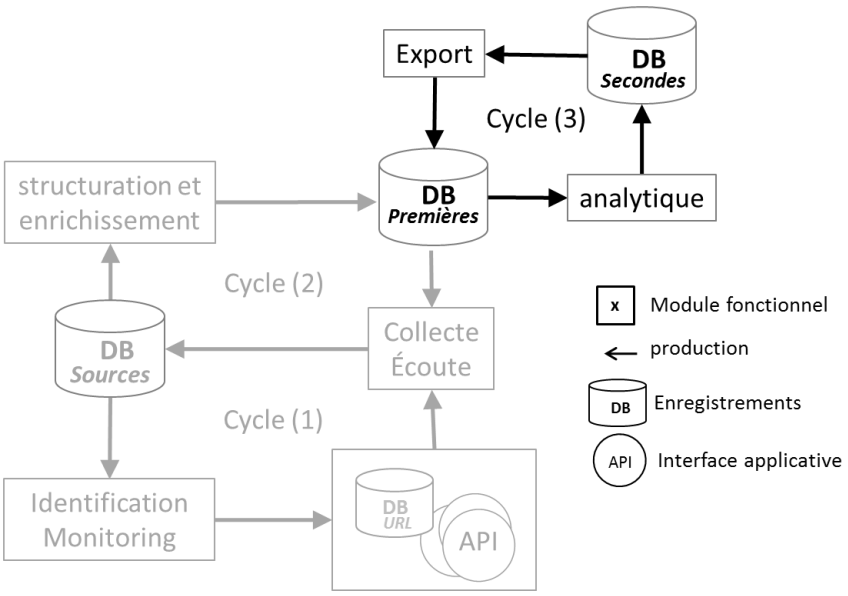

Fig26. Cycle 3

3.3.4. Cycle 4 : supervision et régulations externes

Ce cycle de supervision, qui n'est pas contrôlé par la plateforme, repose sur l'analyse externe des corpus. Il souligne les effets de rétroaction qui peuvent se manifester sur les différentes étapes de configuration, en raison de l'interprétation des résultats ou d'éléments du contexte expérimental qui orientent le cours de la collecte. La prise en compte de ce cycle ne peut s'effectuer que par un horodatage des éléments de configuration des étapes clés d'identification et de *monitoring*. Ces traces permettront ensuite d'intervenir si nécessaire.

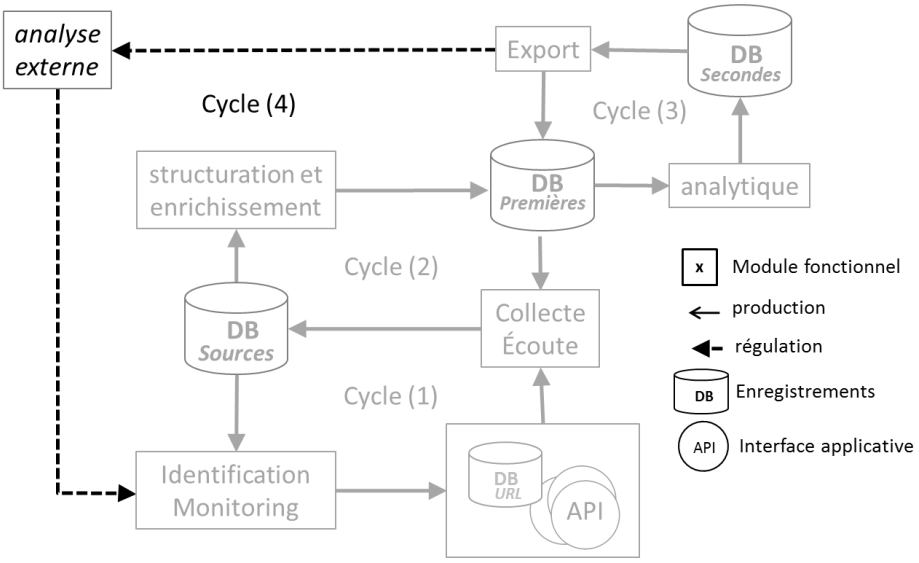

Fig27. Cycle 4

3.4. Visées expérimentales

La modularité fonctionnelle du dispositif permet rapidement d'établir des configurations et des modalités de fonctionnement variées. Ainsi, de manière générale, MEDIASWELL permet la réalisation de *collections* d'entités numériques complexes, élaborées à partir d'éléments récoltés sur l'Internet. Les visées expérimentales sont de ce fait nombreuses et peuvent dépasser le cadre de notre exposé. En ne retenant que les éléments ayant trait aux médiations info-communicationnelles, nous retiendrons les visées suivantes :

- L'analyse des usages et des pratiques médiatisées ;
- L'analyse des interactions interpersonnelles médiées ;
- L'analyse du *bruissement* et de l'*écho*[165] médiatique.

Nous désignons par *bruissement* les informations qui ne font pas encore l'objet d'un traitement médiatique et dont l'importance s'affirmera par la suite. On peut parler de signaux faibles. L'*écho* médiatique traduit les effets de percolation entre les différents média. Il s'agit d'un phénomène d'importation, parfois atténué, provenant d'autres médias et dont on n'a d'ailleurs pas toujours clairement la trace explicite. L'écho se manifeste par des citations, des commentaires et différentes formes de références aux contenus médiatisés ou à la médiatisation elle-même.

4. Synthèse

Les trois dispositifs qui viennent d'être présentés, ont en commun :

- de capter des informations à l'occasion de médiations info-communicationnelles opérées à l'initiative des sujets dans un environnement naturel, non contrôlé ;
- de supporter un ensemble de paramétrages qui assure d'une part, la définition des configurations du dispositif d'observation, d'autre part, le contrôle des conditions expérimentales de la collecte ou de l'usage du dispositif ;
- d'élaborer et de documenter des représentations numériques horodatées complexes en soutien d'une démarche empirique. Cette caractérisation des données, mise en œuvre dès le plus bas niveau d'abstraction (données brutes), est un élément essentiel pour la définition et le contrôle d'un plan de gestion de données (*data management plan - DMP*).

Pour aller plus loin, nous allons dégager les principes de construction et les éléments stables de ces trois plateformes dans la perspective d'une réflexion formelle sur les dispositifs expérimentaux de même type.

[165] C'est essentiellement le registre métaphorique qui fonctionne dans le choix du terme.

4.1. Comparaison des plateformes

4.1.1. Grille d'analyse

Nous pouvons résumer les principales caractéristiques de ces dispositifs dans le tableau récapitulatif ci-après (fig7) dont les facettes se lisent de la manière suivante :

- Trace : unité élémentaire (observable) que collecte le dispositif. Les représentations internes sont élaborées et renseignées à partir de la capture de ces informations ;
- Contexte : cadre(s) socio-technique(s) de l'expérimentation ;
- Mode de collecte : hétéronome ou autonome suivant que la conduite expérimentale nécessite ou pas, l'implication d'au moins une entité agissante (humaine ou non) ;
- Unité d'observation : unité élémentaire d'interprétation dans le contexte de l'expérimentation ;
- Focus : entité faisant l'objet d'un traçage expérimental ;
- Observé : élément calculé à partir des unités d'observation et de modèles dans le cadre d'une collection ;
- Échelle : portée des phénomènes ou ampleur des populations analysées dans le cadre expérimental ;
- Objet d'étude : objet justifiant la mise en œuvre expérimentale de la collection.

	SOPHOCLE	**PLEXUS**	**MEDIASWELL**	
objet d'étude	médiation interaction HM énonciation	pratique web usage web	bruissement et écho médiatique interaction HH médiée pratique médiatisée usage médiatisé	
échelle	micro	meso	macro	
observé	dialogue intervention	session page vue	comportements traits individuels	
focus	**sujet**	**sujet**	ressource	
unité d'observation	intervention	page vue	donnée seconde	
mode de collecte	hétéronome	**autonome**	**autonome**	
contexte	propre	**web**	**web**	Internet
trace	**flux médié, micro-événement**	**flux médié, micro-événement**	donnée première donnée brute	donnée API

Fig28. Tableau synthétique

Les plages grisées mettent en évidence les proximités qu'entretiennent les plateformes entre elles.

4.1.2. Analyse comparée

SOPHOCLE et PLEXUS s'appuient sur des informations produites au plus bas niveau par les outils interactifs[166] utilisés par le *sujet*. Dans les deux cas, la reconstruction de la trace est complexe. Il est nécessaire de réarticuler es flux d'informations avec les micro-événements pour constituer la trace d'usage. Ce point délicat dans le cas du Web a fait l'objet de notre attention dans la conduite expérimentale (Ref.2 p16/343, Ref.6 p93/343). La structuration contemporaine des pages Web, très éclatée et dynamique nous place dans la situation de reconstruire un puzzle d'informations fragmentées.

La granularité informationnelle de la trace numérique ainsi recomposée correspond à ce qui est nécessaire pour reproduire, en dehors de la présence du sujet, l'évolution du poste client à l'identique de ce que les séquences d'actions du sujet avaient produite.

Cette trace d'usage permet l'analyse différée des actions dans une perspective qui va de l'analyse psycho-cognitive à l'analyse de l'usage. Dans le cas de PLEXUS, l'autonomie de la collecte, permet un changement d'échelle ouvrant la voie à l'étude des pratiques info-communicationnelles. Le dimensionnement *meso* indiqué dans le tableau se rapporte au fait qu'une intervention est nécessaire sur l'environnement du sujet, ce qui limite la portée expérimentale à un échantillon d'utilisateur.

La plateforme MEDIASWELL partage avec PLEXUS le contexte d'usage du Web mais dans une perspective différente : l'activité informationnelle du sujet (i.e la consultation de pages Web) ne fait pas l'objet de l'investigation. Dans MEDIASWELL, les pages Web sont associées à l'activité de fouille de données, c'est-à-dire à l'exploration et à l'exploitation de la structure (liens) et des contenus (cf. chapitre 3). Les traces qui découlent de cette activité constituent des données sources dont la structuration (données premières) dépend du projet d'analyse.

En revanche, l'activité info-communicationnelle est étudiée dans le contexte des services de médiation interpersonnelle (ressources) proposés sur Internet. Dans ce cas, l'information délivrée par l'API est conforme au cahier des charges produit par la plateforme de services et ne nécessite pas de reconstruction à ce niveau. Le flux délivré ainsi que la granularité informationnelle des représentations sont contrôlés en amont de leur production. Leurs caractéristiques répondent à des enjeux stratégiques, dans un contexte d'économie de l'information et de la donnée que nous avons évoqué précédemment.

L'échantillonnage ou non du flux et de manière plus générale, les règles de sa production posent la question globale de la *qualité* de la ressource qui n'est pas qu'une généralisation de la *qualité des données*[167] (condition nécessaire) qu'elle contient. Il s'agit également d'une qualité intrinsèque de la ressource, se rapportant à la *pertinence*, à la *représentativité* et à *l'exhaustivité* de l'information contenue au regard d'une analyse. La qualité de Twitter en tant que ressource est une question récurrente de

---

[166] Dans le contexte de l'Internet, ce sont les logiciels clients (navigateurs Web, etc.).

[167] La définition de la qualité des données est très liée à leur exploitation dans des traitements et dans des processus décisionnaires. Dans le premier cas (échantillonnage), ce sont les propriétés calculatoires qui sont mises en évidence et dans le second (non échantillonnage) ce sont les propriétés de représentativité (validité, actualité, etc.) qui sont en jeu.

nos travaux. Celle-ci tient autant des restrictions d'accès dans l'API que des pratiques sociales associées à ce média.

Si l'analyse des flux produits ne permet pas d'évaluer complètement la qualité de la ressource, elle permet néanmoins d'en produire une caractérisation interne qui peut être croisée avec des informations externes. Tenant compte de ces réserves, MEDIASWELL permet d'envisager un changement d'échelle en passant au niveau *macro* analyse. Cette perspective n'implique cependant pas que les résultats d'analyses aient une portée théorique totalisante.

Ainsi, les pratiques info-communicationnelles s'envisagent dans une diversité de contextes d'usages plus étendue au travers de MEDIASWELL mais elles ne restent que partiellement couvertes. De même, les contenus des flux font l'objet des mêmes restrictions pour ce qui concerne l'analyse du *bruissement* ou de *l'écho* médiatique.

4.2. De la trace numérique d'usage à la donnée d'analyse

Pour les trois plateformes, la trace numérique d'usage assure une fonction pivot articulant deux programmes de travaux, dédiés respectivement à la collecte d'information (cadre expérimental) et à la production de données en vue de l'analyse. De SOPHOCLE à MEDIASWELL, les conditions de production de collections d'enregistrement ont évolué en même temps que nos travaux s'écartaient du paradigme cognitif en se rapprochant du paradigme des interactions interpersonnelles.

Les causes de cette évolution tiennent en partie à la transformation du contexte socio-technique global que l'on résumera rapidement par la prédominance croissante de l'Internet parmi les dispositifs de médiations info-communicationnelles. Dans notre activité, le poids socio-technique de l'Internet s'exprime dans un mécanisme antagoniste de normalisation et de différenciation dans la collecte des traces numériques d'usage. La normalisation n'est pas seulement facilitante comme dans le cas de la standardisation des formalismes de données qui en assure l'interopérabilité. Elle est aussi contraignante au travers du développement des *interfaces publiques applicatives* (API) qui adoptent des principes communs : authentification, données *ego*-centrées, limitations en sollicitations des services, limitations en réponse, etc.

La différenciation au contraire nous permet de bénéficier de spécificités de services apportant une complémentarité entre les différentes représentations concurrentes.

La transformation en trois étapes du paradigme de performance que nous avons relevée ne saurait s'expliquer uniquement par l'évolution socio-technique. Nos travaux sont, à leur manière, les témoins d'une autre évolution qui porte sur l'économie numérique et la structuration en marché du Web. Ils en sont les témoins parce que nous nous sommes adapté en suivant l'évolution de l'offre technique correspondant aux méthodes et outils mis à disposition par les acteurs des industries et des marchés du Web eux-mêmes. En accompagnant le mouvement, nous avons pu poursuivre, un temps, un programme indépendant d'études sur les logiques de l'usage. L'utilisation des fichiers de

logs comme moyen n'est pas une concession très forte aux logiques industrielles[168]. En revanche, l'utilisation des flux de traces désormais produits par certains de ces acteurs n'est pas du tout neutre (cf. chapitre 6 §2.). Étudier l'usage d'un service à partir des données produites par ce service sur son usage contient une circularité préjudiciable[169] à la qualité des résultats. Il se trouve que dans le même temps, nos questionnements ont aussi changé de nature. Il ne s'agit plus d'étudier le *comment* une machine ou un service est utilisé mais le *pourquoi.* L'objectif n'est donc plus de comprendre la médiation fonctionnelle qu'assure le service mais sa finalité, c'est-à-dire l'activité qui est visée par l'usager et qui dépasse le service.

Si chacune des plateformes de services se constitue en dispositif par la mise en place d'un système de traces, le changement de perspective que nous abordons fait apparaitre des combinaisons complexes de services organisées par les usagers. Ce sont ces organisations que nous cherchons à étudier. La complexité compense en partie la perte d'autonomie que nous avons consenti vis-à-vis des flux de traces rendus disponibles.

4.2.1. La double contrainte des API

D'un côté, le traçage n'est plus à organiser, il est inhérent au dispositif info-communicationnel qui en organise la mise à disposition. De l'autre, les traces accessibles sont incomplètes et répondent à une stratégie de diffusion informationnelle contrôlée.

Cette double contrainte (*double-bind*) ou injonction paradoxale nous incite à privilégier l'utilisation de ressources à disposition tout en développant des stratégies de contournement destinées à compléter la collecte.

La convergence normative est cependant contrebalancée par la nécessité de différenciation des plateformes de services. Cette différenciation s'exprime dans la richesse et la singularité des attributs accessibles depuis chaque API. Ainsi, les restrictions des interfaces peuvent en partie être compensées par la diversification des ressources collectées. C'est en partie la raison qui nous pousse à organiser des collections d'enregistrements simultanées à partir de plusieurs ressources voire parfois, à partir de plusieurs stratégies de collecte appliquées à la même ressource[170]. Il est cependant difficile d'étendre la connaissance sur une même entité - par exemple le sujet - sans une connaissance externe permettant de relier les identifiants (en général uniques) qui lui sont associés dans chacune des plateformes. Le développement du Web des données tend à lever les difficultés liées à la résolution de la référence au sein du Web. Dans les cas simples, les différentes ressources s'articulent sur les références résolues des entités, ce qui permet de composer une représentation unifiée du sujet.

---

[168] D'autant plus qu'il s'agit de logs de nos serveurs la plupart du temps.

[169] Nous reviendrons au chapitre 6 sur le questionnement épistémologique ouvert par cette absence de neutralité.

[170] Nous avons ainsi démultiplié les écoutes de Twitter (plusieurs collections) en utilisant les deux interfaces API (SEARCH et STREAM) ou en suivant des équations distinctes. Ces collections ont ensuite été fusionnées.

La multiplication des ressources collectées permet également la mise en contexte des données. Il s'agit dans ce cas de produire des métadonnées associées aux représentations afin de rendre compte des conditions de leur production. Nous distinguons alors les traces d'usage des traces contextuelles qui les enrichissent.

Ainsi, dans le cas du suivi d'un événement médiatique dans la *twittosphère* nous travaillons à partir de la ressource Twitter mais aussi à partir de flux RSS provenant de grands médias[171], etc. Dans l'exemple des JO de RIO, la contextualisation est organisée à partir de la temporalité des événements de publication dans les différents espaces médiatiques (Twitter, flux AFP, flux RSS des journaux, etc.).

4.2.2. Documenter les données expérimentales

Un dernier point est également apparu de manière très claire dans l'élaboration de la plateforme MEDIASWELL. Il s'agit de l'enrichissement des données dans le cadre d'un processus documentaire spécifique : chaque étape de la collecte donne lieu à une documentation se rapportant au processus de collecte et d'élaboration des données. Les métadonnées ainsi produites rendent compte de la configuration et du paramétrage de l'ensemble des composants du dispositif expérimental. Ces informations que l'on formalise dans le plan de gestion des données doivent permettre de reproduire les conditions expérimentales de la collecte et d'en maîtriser les variations.

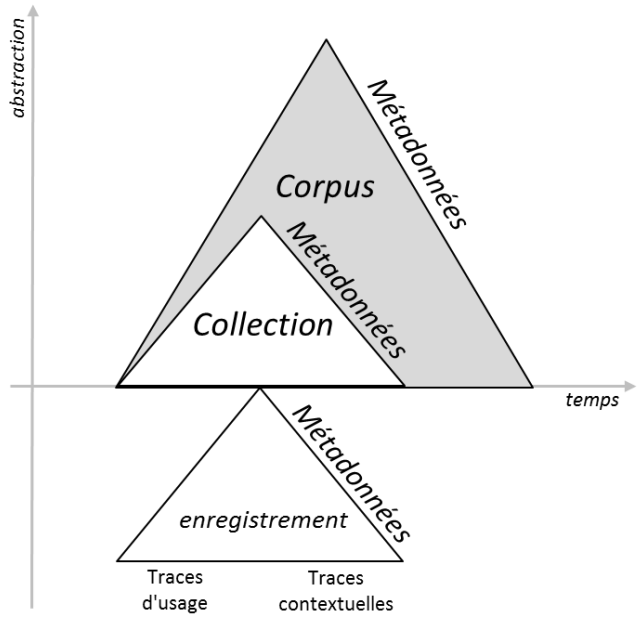

Fig29. Structuration des données

La mise en œuvre de métadonnées, au plus bas niveau de la production de données justifie la réalisation d'un dispositif spécifique tel que MEDIASWELL. Un exemple est présenté en dans le complément consacré au dispositif (ci-après §2.2.2).

[171] Voir par exemple le fil AFP que nous introduisons à partir des jeux olympiques de RIO (2016) et des présidentielles de 2017.

Complément au chapitre 4 : Le dispositif MEDIASWELL

Le texte suivant a été rédigé comme un complément du chapitre précédent. Il s'agit d'un texte technique et indépendant qui décrit plus précisément les choix représentationnels et architecturaux qui ont conduit à la réalisation de la plateforme expérimentale MEDIASWELL. Il nous a cependant paru intéressant de le laisser dans le corps du mémoire, à proximité du chapitre précédent, car la définition des processus et l'organisation des structures de données ne sont pas neutres dans la réalisation de l'objet collection et de sa portée heuristique.

Comme nous l'avons souligné, plusieurs contraintes entrent en ligne de compte dans l'architecture de la plateforme et en constituent le cahier des charges :

(1) La flexibilité attendue d'un dispositif qui doit répondre à des situations et des protocoles expérimentaux variés, incorporant des extensions fonctionnelles à la demande ;
(2) La nécessité de documenter, de produire et de journaliser les traces d'exécution et des états de l'environnement associés à l'expérimentation ;
(3) La spécificité des services de l'Internet et du Web dont on souhaite capturer et analyser les productions et avec lesquels il faut communiquer ;
(4) L'hétérogénéité en nature des données sources que l'on récolte. La dynamique et les volumes des flux de données sources qui doivent être capturés simultanément sans affecter les performances globales de la plateforme et tout particulièrement de la collecte ;
(5) L'enrichissement des données qui passe par l'agrégation de données et l'attribution de catégories à l'issue de traitements et d'analyse ;
(6) La possibilité de mettre en œuvre des analyses au fil de l'eau, dans un délai le plus bref possible permettant d'aborder les problématiques du *juste à temps* ;

Le premier point est un méta principe s'appliquant à l'ensemble des choix architecturaux. La modularité dépend des spécificités des traitements et des contraintes de leur exécution. Le second point conduit à développer un module de supervision. Celui-ci accomplit une méta fonction en interprétant les traces d'exécution de tous les autres modules ainsi qu'en intervenant sur leur programme d'exécution. Cette spécificité fonctionnelle à des conséquences dans l'implémentation des modules. Les quatre points restant nous conduisent à une architecture (Fig30) qui distingue deux autres axes de développement. Le premier est centré sur les données, et concerne les points (3, 4, 5). Le second axe répond aux attentes d'une immédiateté dans l'accès aux résultats de traitements et passe par la mise en œuvre d'un module de visualisation instantané.

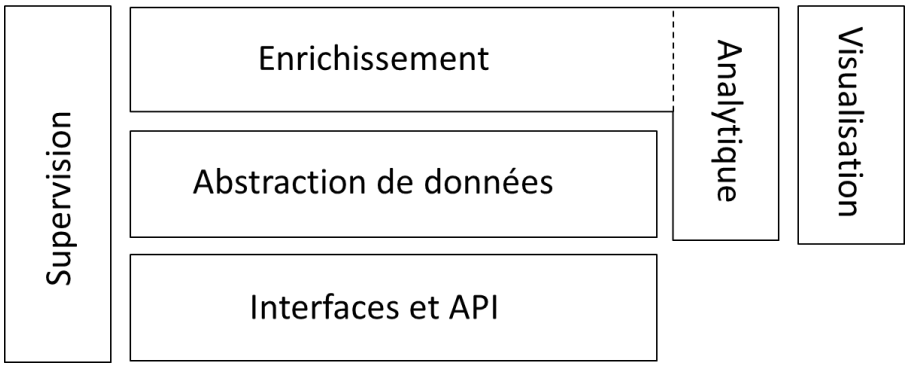

Fig30. Architecture logicielle MEDIASWELL

Nous entendons par abstraction des données, l'ensemble des traitements contribuant en premier lieu à la description et à l'enregistrement des données sources et en second lieu, à la définition des données premières.

La représentation pivot des données est réalisée en JSON[172]. Ce choix s'explique du fait de sa généralisation en tant que format d'échange ; la majorité des API acceptent ce format comme interface. JSON est par ailleurs très proche des objets structurés (dictionnaires) en python. Enfin, le format JSON est particulièrement bien adapté à l'enregistrement des données dans le système de gestion de base de données documentaires qu'est MongoDB et que nous utilisons. C'est pourquoi nous adoptons ce formalisme comme support d'enregistrement et comme vecteur d'échange de données entre les différents modules de la plateforme.

Les principaux modules conversant avec le serveur de base de données sont réalisés à l'aide du module client pymongo[173]. Le module gridfs[174] permet quant à lui de gérer dans MongoDB des enregistrements binaires de grande taille telle que les images et les flux multimédia.

Les données d'entrées sont capturées par les modules spécialisés. Ces données sont exprimées soit dans un dialecte XML (*scraping* Web, moteur de recherche) soit directement dans le format JSON. Les données de type XML issues d'un *scraping* sont conservées en l'état et encapsulées dans un enregistrement JSON. L'objectif est de ne pas effectuer instantanément l'extraction de données si la structure originelle est complexe (cas des pages LinkedIn ou FaceBook par exemple). Dans le cas de structure de données simples non ambiguës, la conversion au format JSON, supportée par de nombreuses librairies, ne pose guère de problèmes.

Une fois les données sources traduites (ou encapsulées) dans le formalisme pivot, des métadonnées de collection sont ajoutées. Celles-ci concernent :

- l'horodatage de la capture pour établir une chronologie cohérente dans le système ;
- une identification de collection ;
- les spécificités du contexte d'interaction avec l'API ou la ressource Web ;

[172] http://www.json.org/jsonfr.html

[173] https://api.mongodb.org/python/current/

[174] http://api.mongodb.org/python/current/api/gridfs/

- des indications de traitements pour certains modules.

À la suite de cet estampillage, les données sont enregistrées dans la base de données. Lorsque celles-ci ont vocation à être unique, les doublons sont comptabilisés (et rejetés). Les cas nécessitant une unification des données sont relativement rares pour les données sources.

1. Réalisation informatique

1.1. Implémentation de classes de processus

Chaque module décrit un ensemble de fonctionnalités liées à l'accomplissement d'une classe de traitements suivant une logique de programmation orientée objet. L'implantation python repose sur le module YapDi[175] qui permet l'instanciation d'une classe en processus. Cette solution permet de déployer autant d'instances d'un module que nécessaire sur autant de machines différentes.

Chacune des instances de module constitue un agent (processus) autonome dont l'activité est contrôlée de trois manières différentes :

- Par le paramétrage initial du processus, décrit dans un fichier de configuration qui établit certains éléments d'interaction avec le système de représentation des données et avec d'autres agents ;
- Par le système de représentation des données partagées qui comporte des métadonnées utilisées dans le contrôle d'activité ;
- Par un module de supervision global qui contrôle le programme d'exécution des processus et vérifie leur état de fonctionnement.

Tous les modules de traitements produisent des traces d'exécutions associées aux processus qu'ils accomplissent. Ces traces intelligibles sont horodatées et journalisées (*logs*). Des indicateurs d'état sont également enregistrés dans la base de données pour permettre un *monitoring* supervisé.

1.2. Module de supervision

La supervision est prise en charge partiellement par un module spécifique qui scrute en permanence, suivant une boucle d'attente, l'état de chacun des modules et veille à l'exécution d'événements programmés.

L'état des modules est évalué à partir des informations que peuvent communiquer les processus instanciés. La réponse de ceux-ci est une indication qui est confortée par l'analyse de leurs productivités, via les enregistrements dans les collections. Si l'un des processus semble inactif, ne répond plus ou est mort, il est automatiquement ré-instancié. Si la réactivation est impossible un système d'alertes par *email* est possible.

Plus globalement, le cycle de vie ou d'activité des processus est réglé par un programme global décrivant les événements d'activation associés aux modules. Les événements sont déclenchés à

[175] https://github.com/kasun/YapDi

partir de la description d'une plage d'activation qui peut être répétée ou non dans un cycle (horaire, quotidien, etc.). On associe à chacune de ces plages une liste de module actifs.

Le programme d'activation est stocké dans le fichier de configuration du module de supervision. La description d'un événement d'activation suit la syntaxe suivante:

<horodate1> <horodate2> <cycle> | [<module>]+

Ainsi, l'événement suivant :

05/02/16-00:00:00 08/02/16-00:00:00 jamais | TwitterAPI Archiver NLPParser URL Media

Provoque l'activation programmée du 5 février à 00h au 8 février 00h sans répétition pour cinq processus réalisant respectivement les fonctionnalités de capture, d'archivage, d'analyse TAL, de complétion d'URL et de téléchargement iconographique pour une collection du flux de Twitter réalisée à l'occasion de la journée du sport féminin. Cette programmation collective des cinq processus est équivalente à la programmation individuelle de chacun d'entre eux.

Il est possible de déclarer plusieurs plages de fonctionnements et constituer ainsi des programmes complexes (plusieurs activations journalières par exemple) répétées ou non un certain nombre de fois.

1.3. File d'attente

Pour permettre à la plateforme d'absorber un flux de données d'entrées important sans altérer les performances de traitement, tout particulièrement dans une perspective de *juste à temps*, nous utilisons un dispositif de file d'attente.

Le dispositif de file d'attente est pris en charge par un serveur *RabbitMQ*[176] qui permet d'administrer simultanément plusieurs files d'attentes et d'effectuer sur celles-ci des opérations d'interclassement ou de dédoublement. Ce serveur permet également de synchroniser des processus en leur délivrant simultanément la donnée de tête de file.

L'utilisation de file d'attente permet d'établir un tampon et désynchroniser des traitements consécutifs qui n'ont pas la même célérité. Dans le cas du flux de Twitter, le module[177] d'interface avec son API et le module d'archivage communiquent à travers une file d'attente qui absorbe les flux d'entrée lorsque ceux-ci sont très importants. Cette solution maintient un temps quasi constant en acquisition et lisse la charge pour le serveur de données.

Le mécanisme de file permet d'organiser conjointement une activité longue de constitution d'archives fortement enrichies avec une activité de visualisation instantanée restituant dans leur dynamique des phénomènes saisis au fil de l'eau, comme par exemple l'évolution du flux d'un *hashtag* sur Twitter.

---

[176] Basé sur la technologie AMQP (Advanced Message Queuing Protocol), ce serveur réalise un bus de données asynchrone http://www.rabbitmq.com/documentation.html

[177] Les modules nécessitant l'accès aux données en file d'attente sont déclarés comme sous-classes de clients *amqplib*. https://pypi.python.org/pypi/amqplib.

1.4. Moniteur de données

Un module spécifique qualifié de moniteur de données a pour objectif de constituer un corpus de données temporelles à partir des données de collections ou des données de corpus. La dimension temporelle est soit exprimée dans la donnée (par exemple horaire de publication), soit issue d'un processus de traitement dans MEDIASWELL, comme par exemple l'archivage.

La grille temporelle qui est associée à ces indications fait référence soit au temps horaire soit à une chronologie permettant de distinguer différents états successifs.

Le moniteur permet ainsi de composer des vues diachroniques ou synchroniques sur les données qui alimentent des Web services chargés de la visualisation[178].

1.5. Modules Interfaces Web et API

Communiquer avec les ressources d'Internet passe par la mise en œuvre de protocoles dont HTTP est l'un des plus connus. La majorité des ressources informationnelles que nous avons explorées jusqu'à présent le sont via ce protocole permettant la circulation de documents dans des formats de type HTML ou XML.

Ce protocole s'il donne un accès public à de nombreuses ressources ne permet généralement pas d'accéder aux représentations informationnelles les plus riches. C'est par exemple le cas avec les pages publiques des réseaux sociaux. Dans le cas de Twitter, la page Web associée à un compte utilisateur donne effectivement accès aux contenus des messages qu'il canalise. En revanche, la page ne fournit aucune métadonnée sur les émetteurs des messages.

Pour accéder aux contenus les plus complets il est préférable, lorsque cela est possible, de passer par les interfaces publiques d'accès aux données que sont les API. Cependant, suivant la nature du projet, il est parfois nécessaire de conserver une approche Web dégagées des contraintes contractuelles et techniques qu'imposent les fournisseurs d'accès.

1.5.1. Extraction d'information des pages Web

La fouille de données du Web consiste à exploiter l'information structurelle contenue dans les liens ou l'information de contenue inscrite dans les pages. La distinction peut apparaitre ténue car l'analyse structurelle nécessite d'accéder au contenu et réciproquement. Cependant, pour des raisons d'efficacité, les spécificités de l'une et de l'autre conduisent à la réalisation de modules distincts.

Le module arpenteur (*Web Spider*) est spécialisé dans l'analyse structurelle du Web. Il est conçu pour explorer les liens extraits des pages parcoures suivant différentes hypothèses et heuristiques qui tiennent à la nature des liens, le degré du nœud atteint (page), la longueur du trajet accompli, etc. Ce module ne met pas réellement en œuvre une analyse de contenu et s'appuie plutôt sur les métadonnées. Le module est réalisé à partir du *framework open source* Scrapy[179]. Cette technologie est

[178] Des expérimentations sur la visualisation sont en cours et constituent un pan important des développements à venir.

[179] Plus précisément le spider du module : https://scrapy.org/

mise en œuvre dans le cadre de la collaboration régionale issue de l'action de recherche ARC6 (cf. Ch6 §3.2.2).

Le *Web scraping* consiste, à partir d'une URL de page Web, à extraire les informations pertinentes qu'elle contient et qui constituent alors la trace documentaire. Ces pages étant le plus souvent dynamiques (technologie ajax), on ne peut pas analyser le code source de la page sans exécution des scripts (java scripts) qui lui sont associés. Il est alors nécessaire de passer par sa représentation DOM[180] produite par un moteur de navigateur pour intégrer les différentes transformations nécessaires à la complétude et à l'intelligibilité du document restitué.

L'extraction des contenus des pages Web est réalisée par deux modules distincts.

Le premier module (*Web Crawler*) a pour objectif de produire et d'enregistrer la représentation DOM associée à une URL. Ce module s'appuie sur Selenium[181] qui permet de contrôler différents navigateurs (FireFox, Chrome, etc.) et leur moteur de restitution graphique (Gecko par exemple). Le comportement graphique du navigateur est simulé[182]. Le contenu XHTML extrait du navigateur (page complète) est ensuite enregistré en base de données dans un format compressé (cf. §3.6).

Le second module (XML scraper) procède à l'extraction effective des données à partir du parcours de l'arborescence XML (cf. § 4.6.1)

1.5.2. Interfaces Applicatives de Programmation (API)

L'accès par les interfaces applicatives (API) impose de mettre en œuvre de modules clients spécifiques qui réalisent les fonctions de communication avec les services concernés suivant leur protocole. En général, ces modules spécialisés existent déjà. Il s'agit alors de réaliser un adaptateur (*wrapper*) qui adapte le client aux logiques et spécificité du système MEDIASWELL.

S'agissant de modules de communication, une structure de *wrapper* fortement générique est adoptée pour limiter les développements spécialisés et leurs coûts. Cette simplicité de mise en œuvre nous permet de développer des modules distincts spécialisant des modalités d'interaction différentes avec les services du Web. C'est par exemple le cas avec les modalités *search* et *stream* de l'API Twitter mises en œuvre dans le même client (*Tweepy*[183]).

Outre l'adaptateur pour la communication avec des services tels que Twitter ou YouTube, nous avons développé des adaptateurs pour différents moteurs de recherche (Google, Bing) Le module Bing est le principal que nous utilisons. La raison tient à la qualité de l'interface qui accompagne une stratégie commerciale claire de la part de Microsoft[184]. Le module Bing est utilisable suivant les

---

[180] Document Object Model – standard du W3C permet l'analyse indépendamment du format HTML, XML, XHTML du document.

[181] http://docs.seleniumhq.org/

[182] https://pypi.python.org/pypi/PyVirtualDisplay

[183] https://github.com/tweepy/tweepy

[184] Voir le site http://datamarket.azure.com.

deux modalités décrites précédemment (cf. 2.1.1.) pour extraire les URL ou les résumés qui leur sont associés.

1.6. Modules liés à l'enrichissement et complétude des données

1.6.1. *Scraping* XML

Le module (XML scraper) opère sur les enregistrements comportant des fragments XML pour en extraire les éléments souhaités. La localisation des contenus est réalisée grâce à la bibliothèque *BeautifulSoup*[185] (bs4). L'extraction est conçue en deux temps, pour des raisons d'efficacité.

Elle met en œuvre un automate dont le paramétrage est spécifié au moyen d'une grammaire. Cette grammaire permet de décrire les chemins (*paths*) conduisant aux sous-arbres XML contenant l'information visée. La grammaire permet de signifier plusieurs chemins alternatifs pour rendre compte de la diversité des formes structurelles[186] (variabilité des pages). En général quelques cas suffisent pour couvrir l'intégralité des variations rencontrées.

1.6.2. Résolveur d'URL

Ce module est spécialisé dans le traitement des URL capturées. La généralisation des formes abrégées d'URL (*tiny URL*) implique de résoudre de l'URL pour en trouver la forme expansée. Nous avons pu constater que cette étape de traitement doit être effectuée très peu de temps après la collecte de l'URL afin de ne pas perdre la référence exacte. La dégradation est en effet rapide : environ 20% des références sont perdues après un mois.

La résolution de l'URL nécessite de contacter la ressource associée à l'URL (raccourcie ou non) pour obtenir en réponse l'URL réelle. Cette contrainte allonge le temps de traitement ce qui rend nécessaire l'existence de ce module opérant de manière autonome sur la base de données. Outre la résolution, ce module normalise l'encodage de l'URL. Celle-ci peut en effet être décrite au format HTML, répondre à la norme IDNA[187] ou encoder des caractères accentués. Enfin, les problèmes de sur-encodage (ASCII, UTF8) peuvent aussi être rencontrés.

L'URL ayant été normalisée, des traitements opérés sur les noms de domaines ou sur les chemins peuvent être envisagés.

---

[185] http://www.crummy.com/software/BeautifulSoup/bs4/doc/

[186] Le formalisme de référence pour décrire la grammaire est le formalisme JSON. Les expressions grammaticales peuvent contenir des expressions régulières. Par exemple : {"prof": '#<div id="experience-[\d]*-view>'} signifie que la localisation de la donnée brute associée à l'élément structurel "prof" est associée à une balise de subdivision ( ) du bloc et que cette balise a pour identifiant ('id') une valeur littérale dont le préfixe contient "experience-" suivi d'une valeur numérique suivi du suffixe "-view" (ex : "experience-1234-view" est une expression valide.)

[187]Voir à ce sujet https://www.afnic.fr/medias/documents/afnic-idn-specifications-techniques.pdf

1.6.3. Téléchargement de documents multimédia

Les composants multimédias (iconographiques ou audiovisuels) jouent un rôle de plus en plus important dans la communication instantanée. Introduit en 2014 sur Twitter, les images *Tweetées*[188] ont modifié les logiques de publication. L'introduction imminente des composants audiovisuels[189] dans Twitter va engendrer de nouvelles pratiques et modalités conversationnelles qu'il nous faut anticiper et étudier.

Depuis 2014, un module spécialisé permet de constituer des collections de documents référencés et associés à la collection de données primaires. Pour des raisons semblables à la résolution des URL, la résolution des références aux composants multimédias doit être effectuée au plus tôt de même que le téléchargement de la ressource.

Le téléchargement des composants multimédias implique une durée de transfert non négligeable justifiant de l'existence d'un module autonome opérant à partir de la base de données. Les contenus téléchargés sont enregistrés dans la base mongodb[190]. Leur visualisation est possible dans les clients mongodb des fonctionnalités graphiques ont été réalisées pour explorer sous cet angle les collections.

1.6.4. Traitement Automatisé de la Langue

Un module TAL assure la fonction d'interface avec les services Web permettant d'accéder aux Web services d'analyse linguistique (*Tagging* lexical, segmentation, etc.) mis en œuvre par le LIDILEM dans le cadre de collaborations191. Ce module permet la normalisation des formes textuelles et l'enrichissement des métadonnées linguistiques des enregistrements primaires.

1.6.5. Traitement Sémantique

Ce module a pour fonction est d'assurer l'enrichissement sémantique des représentations associées aux données textuelles et principalement aux entités nommées ; Le module d'analyse sémantique intervient sur les données premières après que l'analyse linguistique ait été réalisée. Le module linguistique produit une représentation augmentée du message originel. En particulier, certaines entités nommées sont reproduites sous une forme littérale complète et normalisée. À partir de cet enregistrement, le module sémantique peut interroger le service Web OpenCalais pour valider ou compléter la catégorisation sémantique de ces entités. Il peut alors, en fonction des catégories et des besoins analytiques, interroger le service Web DBPédia pour enrichir (sémantiquement) ou documenter ces entités et de là l'enregistrement.

---

[188] Il s'agit d'images et plus généralement de contenus audiovisuels. À l'origine seule l'URL était transmise dans le corps du texte. Le protocole de Twitter permet depuis 2014 de visualiser instantanément ces contenus dans les interfaces graphiques. Ainsi le *tweet* s'aligne dans sa forme avec les billets présents sur les murs de Facebook.

[189] Twitter a acquis la start-up *Periscope* en mars 2015 et associera les vidéos réalisées au flux de Twitter à partir de mars 2016.

[190] https://www.mongodb.org/

[191] Projets TRI-Elec (Présidentielle 2012) ; PEPS Humain (Européenne 2014) ; ANR RJS-MéDIS (RIO 2016).

2. Exemple de configurations

Suivant le principe de modularité, l'adaptation de la plateforme aux conditions expérimentales revient à : identifier les modules de traitements actifs, définir leur paramétrage (fichier de configuration) ainsi que les conditions de leur activation (script de *monitoring*).

2.1. Identification des comptes Twitter (Cycle 1)

La configuration suivante est utilisée pour identifier les outils de communication des athlètes français engagés dans les jeux olympiques[192]. Dans cet exemple nous nous intéressons aux comptes Twitter. Mécaniser cette extraction évite un fastidieux travail de collecte manuel qui ne peut être validé que tardivement par rapport à l'événement que l'on veut suivre. Le site officiel de l'équipe de France[193] maintien à jour la liste des athlètes sélectionnés ainsi que des fiches individuelles.

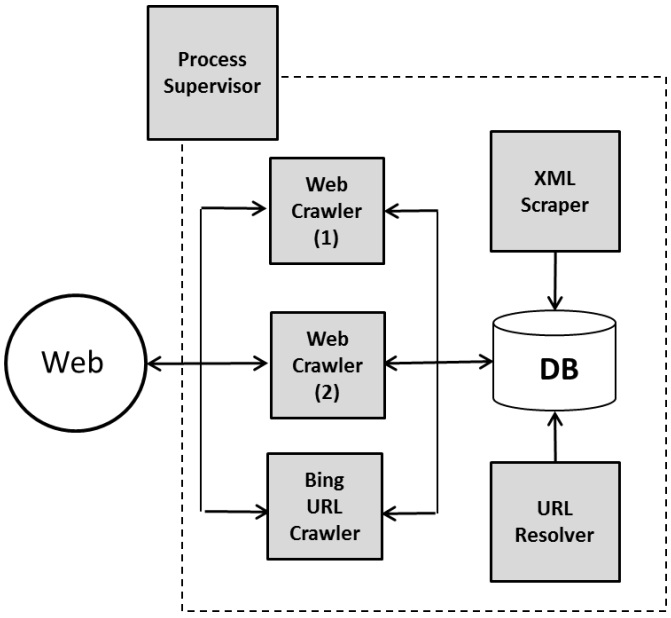

Fig31. Organisation fonctionnelle des modules pour l'identification de sources.

Les modules sont représentés par des boîtes grises. L'élément DB correspond à la base de données dans laquelle sont enregistrées les données (URL, sources, premières, secondes).

Dans le cas présent, la méthode consiste à fournir l'URL (seed URL) du site référent en entrée du premier module WebCrawler (1). Celui-ci charge la page et parcours sa structure pour extraire les URL des fiches individuelles des athlètes sélectionnés. Celles-ci sont alors stockées dans une table d'URL[194]. Le second module WebCrawler (2) utilise ces URL sélectionnées pour accéder aux pages individuelles et les enregistrer en tant que données sources. À partir de ces enregistrements[195], le

[192] Testé lors de SOCHI2014 sera systématisé à l'occasion de RIO2016

[193] http://espritbleu.franceolympique.com/espritbleu/equipe-de-france.php

[194] Les URL étant spécifiques, il n'est pas nécessaire d'activer un module d'expansion.

[195] L'évolution du Web sémantique devrait permettre d'exploiter davantage cette fiche en tant que telle.

module WebScraper construit une représentation première de l'individu contenant son nom, sa discipline, etc.

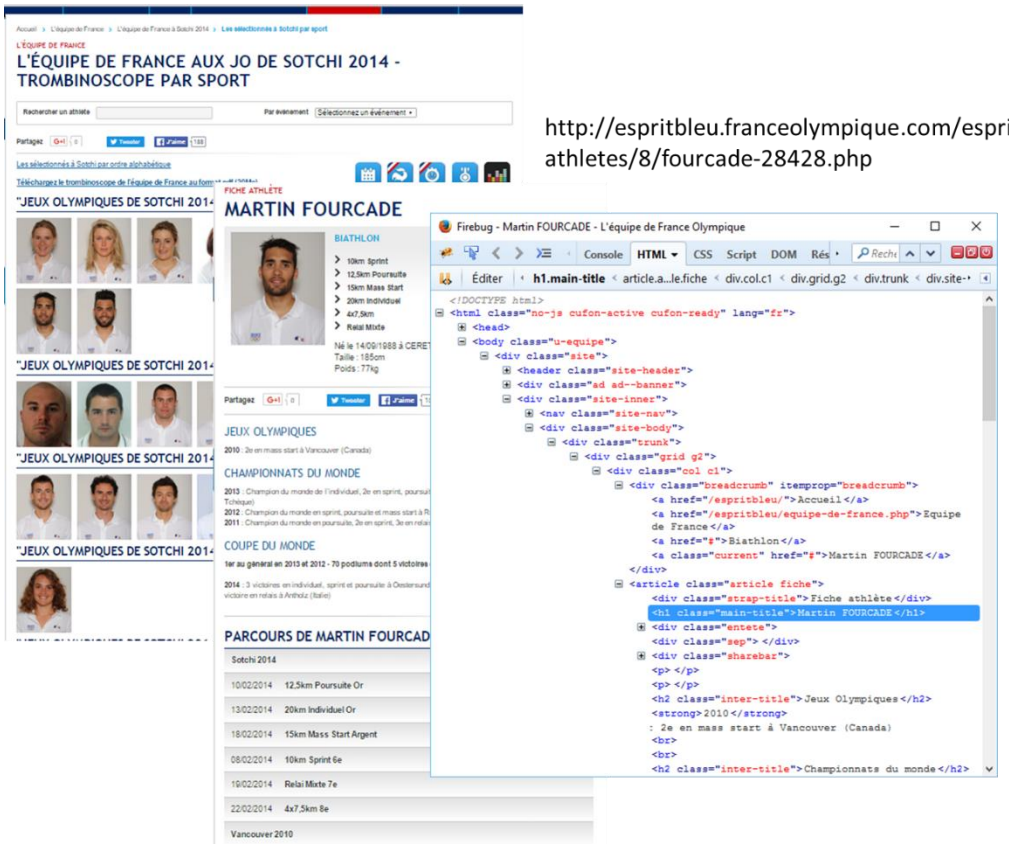

Fig32. Web Scraping appliqué à la recherche d'information
sur les athlètes sélectionnés au JO d'hiver (Sochi 2014).

Pour chaque athlète, les informations ainsi extraites sont associées à une requête qui va être soumise au moteur de recherche Bing. Cette requête `<prénom nom> AND site:twitter.com`, produit une page de résultats privilégiant les informations Twitter.

Parmi les scénarios possibles, nous privilégions celui qui consiste à enregistrer les URL de la première page de résultats qui satisfont un schéma lexical et syntaxique particulier[196].

Cette méthode permet de collecter les URL de comptes associés à des identités qui peuvent être associées à plusieurs profils ou a plusieurs individus (ex "Martin Fourcade"). Une équation plus restrictive : `<prénom nom> AND site:twitter.com AND biathlon`, n'identifiant que l'athlète aurait pu être envisagée. Les deux stratégies de requête peuvent être menées simultanément, l'une apportant confirmation de l'autre. Pour alléger la tâche de contrôle, nous avons testé différentes hypothèses de confirmation par les contenus (pages des comptes). Mais dans tous les cas, l'intervention humaine est indispensable.

---

[196] Cette restriction est opérée à l'aide d'une expression régulière qui isole un segment du chemin (*path*). Cette méthode permet de sélectionner la référence au compte Twitter à partir de *tweet* (du compte) ayant été indexé par le moteur de recherche : https://twitter.com/martinfkde/status/1111111 produit

https://twitter.com/martinfkde

À partir de la validation des informations des comptes Twitter, la collecte des Tweets publiés par les athlètes (TwitterCrawler) peut être mise en place.

2.2. Collecte de flux de Tweets, un exemple : "sport au féminin"

Le week-end du 6-7 février 2016 a été associé au sport féminin dans plusieurs pays dont la France. Il a été médiatisé sous l'intitulé des *4 saisons du sport féminin* à l'initiative de Nathalie Sonnac membre du CSA et présidente de la commission sport. Il a été promu au plan national en partenariat avec le ministère des affaires sociales, de la santé et des droits des femmes, le ministère de la ville, de la jeunesse et des sports et le comité national olympique (CNOSF)[197]. France télévision a consacré plusieurs programmes à cet événement sportif et médiatique[198].

La collection *sport au féminin* associé à cet événement vu depuis Twitter a été réalisée dans le cadre de l'ANR RSJ-MéDiS[199]. L'objectif de cette collection est d'évaluer si un espace de discours particulier peut s'organiser dans Twitter à l'occasion d'un événement médiatique animé par FranceTV sur la durée du week-end.

La constitution de cette collection minimaliste (non contextualisée) a été justifiée comme pré-expérimentation en vue de tester le dispositif et d'évaluer la portée de cet événement médiatisé. En particulier, nous étions intéressé par l'éventualité de repérer (ou non) des éléments d'un discours caractérisé par l'opposition masculin/féminin dans le domaine du sport.

2.2.1. Architecture fonctionnelle

Le dispositif est donc centré sur la collecte de données de Twitter via l'API Stream qui délivre un flux continu à partir d'un filtre. Le filtre choisi est en l'occurrence l'équivalent de la requête logique : sport AND féminin OR sportféminin OR francetv AND sport OR Francetvsport.

Ce filtre réunit quatre flux de Tweets pouvant se recouvrir. L'expression "sport AND féminin" canalise un flux de tweets contenant les deux mots. Le mot "féminin" est reconnu en tant que partie d'un mot plus long (dont en particulier #féminin ou @féminin), accentué ou non, sans considération pour la casse.

L'expression complète réunit ainsi deux collections, l'une liée à l'événement lui-même, l'autre au média (FranceTV) impliqué dans sa promotion médiatique.

---

[197] URL (consultée le 03/07/2016) http://www.csa.fr/%20csa/Radio/Les-4-saisons-du-sport-feminin-Quatre-temps-forts-pour-la-feminisation-du-sport-dans-la-societe-francaise

[198] URL (consultée le 03/07/2016) : http://www.francetelevisions.fr/node/1226.

[199] Responsabilité sociale des journalistes : Média, Diversité, Sport. ANR-15-CE26-0006-01. Projet consacré à l'expression de la diversité dans le discours journalistique ou en lien avec celui-ci.

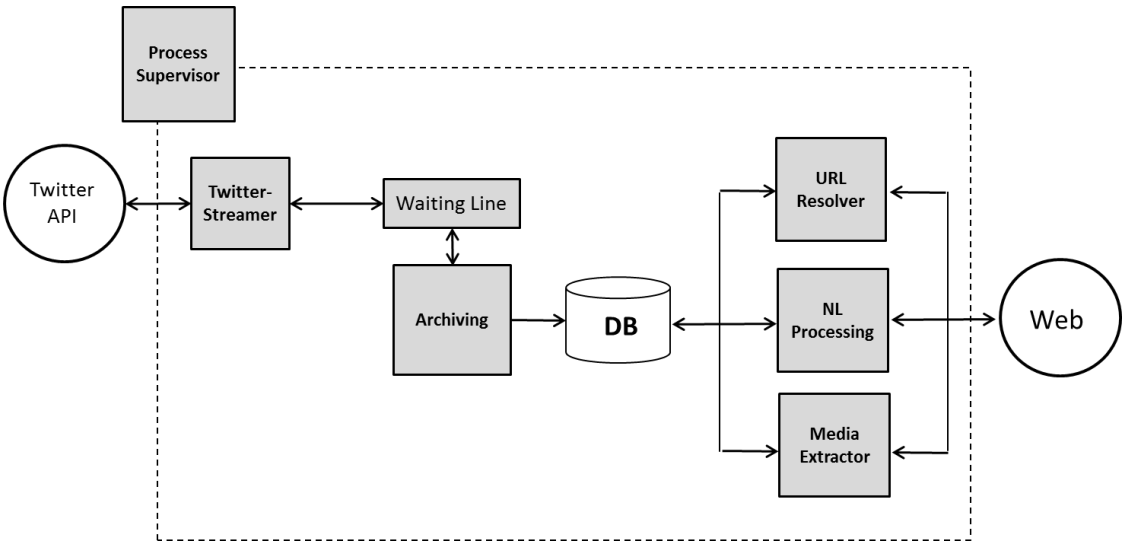

Fig33. Organisation fonctionnelle des modules pour la collecte
de type événementiel médiatique sur Twitter.

Le module TwitterStreamer paramétré avec l'équation précédente est à l'écoute de l'API Twitter et transmet les Tweets dans la file d'attente. Celle-ci est épuisée par le module d'archivage qui construit la collection de messages tout en constituant simultanément un éclaté de la structure du message isolant les données premières relatives aux abonnés, aux *hashtags*, aux URL, aux composants multimédias. Cet éclatement est opéré afin de favoriser les analyses qui suivront.

Les trois modules URL Resolver, Media Extractor et NL Processing contribuent à l'enrichissement asynchrone des données premières.

2.2.2. Plan de gestion de données : "sport au féminin"

Le plan de gestion de données associé à cette collection est simplifié ; raison pour laquelle nous présentons dans ce mémoire la partie descriptive des données sous forme synthétique.

1	EXPERIMENTATION					
2		Projet				
3			désignation	RSJ-MéDIS		
4			financement	ANR		
5			Date	2016-2019		
6			Ref	Tâche T3		
7		Descriptif				
8			Nature	collection événementiel médiatique		
9			Objet	4 saisons du sport féminin		
10			Début	6/2/16 0:00		
11			Fin	8/2/16 0:00		
12			Ressources			
13				Twitter		
14					Cycle	Permanent
15					Source	API
16						désignation STREAM
17						Modalité Oauth
18						Equation sport féminin,sportféminin,4saisonsportfem,francetv sport,francetvsport
19						Restriction Lang=fr

Fig34. Caractérisation globale de la ressource : Twitter.

Cette partie de présentation du projet indique que Twitter est interrogé au travers de l'interface STREAM suivant l'authentification Oauth. Le flux observé est restreint aux publications dont la langue (calculée) est "fr" et répondant au filtre caractérisé par l'équation suivant le formalisme de requête.

```
Modules
    Twitter Streamer
        Monitoring  oui
        Début       5/2/16 10:00
        Fin         8/2/16 0:00
        Interface   python.2.7-Tweepy
        input       API
        output
            Format      JSON
            API data
                data            Status
                documentation   https://dev.twitter.com/overview/api/tweets
            Metadata
                {'collection' :  ['sport féminin','sportféminin','4saisonsportfem','francetv sport','francetvsport'],
                 'source' : 'streamer',
                 'nlp_status' : 0,
                 'citation' : False,
                 'random' : 0,
                 'event' : False}
    Archiver
        Monitoring  oui
        Début       5/2/16 10:00
        Fin         8/2/16 0:00
        Interface   python.2.7-pymongo
        input       Twitter Streamer
        output
            DataRecord
                Collection      statuses
                Format          JSON
                record
                    data
                        Status
                    Metadata
                        {'archived_at' : datetime,
                         'in_arc' : +1,
                         ' nlp_status' : 0}
            DataRecord
                Collection      users
                Format          JSON
                record
                    data   Status['user']
                    Metadata
                        {'archived_at' : datetime,
                        Tweet_id : bigint
                         'in_arc' : +1,
                         ' nlp_status' : 0}
            DataRecord
                Collection      hashtags
                Format          JSON
                record
                    data
                        Status['entities']['hashtags']* + Status['retweeted_status']['entities']['hashtags']*
                    Metadata
                        {'archived_at' : datetime,
                        Tweet_id : bigint
                         'in_arc' : +1}
            DataRecord
                Collection      urls
                Format          JSON
                record
                    data
                        Status['entities']['urls']* + Status['retweeted_status']['entities']['urls']*
                    Metadata
                        {'archived_at' : datetime,
                        Tweet_id : bigint
                         'in_arc' : +1,
                        'expanded' : 0}
            DataRecord
                Collection      media
                Format          JSON
                record
                    data
                        Status['entities']['media']* + Status['retweeted_status']['entities']['media']*
                    Metadata
                        {'archived_at' : datetime,
                        Tweet_id : bigint
                         'in_arc' : +1,
                        'extracted' : 0}
```

Fig35. Structures de données mises en œuvre dans la capture et l'archivage d'un tweet.

La représentation précédente traduit l'effet de l'activation des modules sur la structure de données. La donnée brute est décrite à partir de l'URL (ligne 12). À l'issue de la capture par le module *TwitterStreamer* (ligne 2), les métadonnées (ligne 14) sont ajoutées avant la mise en file d'attente de la représentation.

Cette représentation est ensuite prise en charge par le module *Archiver* (ligne20) qui va alors produire plusieurs représentations (data records), associées aux : comptes (ligne 37), *hashtags* (ligne 47), URL (ligne 57), média (ligne 68). Pour chacun de ces enregistrements, des métadonnées sont ajoutées (date d'enregistrement, compteur d'occurrence, statut d'extraction ou de traitement linguistique). Ces représentations sont coordonnées à l'aide de la référence *Tweet_id* qui est l'identifiant interne unique du tweet dans le système d'information de Twitter.

Ces métadonnées permettent aux modules spécialisés comme par exemple le *resolver* d'URL (ci-après) de repérer les enregistrements à traiter et d'accomplir leurs traitements.

```
1  Modules
2      URL Resolver
3          Monitoring  oui
4          Début       5/2/16 10:00
5          Fin         8/2/16 0:00
6          Interface   python.2.7-pymongo
7          output
8                      DataRecord
9                          Collection      urls
10                         Format          JSON
11                         record
12                                         data
13                                             url.record
14                                         adding
15                                             {'netloc' : string,
16                                             'expanded_url : 'string'}
17                                         Metadata
18                                             { 'code' :  int,
19                                             expanted_at' : datetime,
20                                             'expanded' : 1}
```

Fig36. Effets du module URLResolver sur la structure de données de la collection urls.

Le module de résolution d'URL s'active sur l'ensemble des enregistrements qui n'ont pas été expansés. Lorsqu'une URL est traitée, les métadonnées traduisent l'accomplissement de l'expansion (ligne 17) et l'enregistrement est étendu (ligne 14).

CHAPITRE 5
Aspects méthodologiques de la collecte de traces d'usage

« *L'information ne devient information*
que par rapport à une computation,
et n'est sinon qu'une marque ou une trace.»
(Morin, 1986, p38)

En conclusion du chapitre 3 consacré aux traces d'usage, nous avons évoqué l'existence de travaux en Sciences de l'information et de la communication (SIC) prenant comme objet d'étude le traçage numérique. Pour une partie de ces recherches d'inspiration sémiologique, les traces prennent place dans un système d'inscriptions et de signes qui répond à une logique culturelle et sociale. Pour une autre, les traces s'inscrivent dans un régime de production de sens dans lequel l'identité personnelle du sujet est adressée et manipulée. Le concept de trace d'usage unifie les deux perspectives. Cette perspective, tout aussi intéressante qu'elle soit, n'est pas approfondie dans ce mémoire.

Suivant la perspective numérique que nous poursuivons, la trace d'usage est une donnée informatique première à partir de laquelle s'élabore dans un dispositif, une représentation computationnelle associée à un phénomène ou un objet, agissant ou agi. Par hypothèse, les objectifs de calcul sont plus étendus que le suivi ou la caractérisation personnelle.

Les problématiques que nous avons soulevées dans le chapitre précédent portent sur l'instrumentation expérimentale nécessaire pour s'approprier des données issues de ces représentations et les enrichir au sein de *collections* exploitables à des fins de recherche. C'est de la manière d'atteindre ces fins qu'il est désormais question.

Dans nos travaux, la collecte de traces numériques d'usage ne peut être un objectif scientifique valide qu'en conjonction d'objectifs analytiques assignés. En reprenant une définition classique de la démarche scientifique, l'analyse de données fait suite à un choix méthodologique effectué dans un cadre théorique, lui-même établi relativement à un objet d'étude. Mais cette démarche ne peut se mettre en place que lorsque les méthodes choisies sont éprouvées et le contexte de leur utilisation maitrisé. Or, les projets de recherche que nous avons conduit n'ont jamais rempli de telles conditions. En effet, les objets de recherche que nous privilégions s'inscrivent le plus souvent dans des contextes socio-techniques mouvants, peu explorés, dans lesquels les méthodologies classiques d'investigation ne peuvent pas s'appliquer ou alors seulement avec un faible bénéfice.

Ces contraintes nous ont incité à privilégier l'approche empirique comme cadre de production scientifique. Cette approche invite à une plus grande souplesse méthodologique. Sans rejeter la

rigueur scientifique, elle nous permet d'explorer autrement ces contextes, d'éprouver des méthodes alternatives ou complémentaires. La constitution de collections s'inscrit dans cette perspective. Nous n'avons généralement pas les moyens de fixer *ex ante* la pertinence d'une collection. Celle-ci ne peut s'élaborer que progressivement, par tâtonnements et ajustements, de manière incrémentale dans un va-et-vient entre élaboration de la collection et exploration des données (pré-analyse). Pour nous accorder pleinement avec la démarche empirique, nous considérons que le processus d'élaboration de collections fait partie d'un processus d'*observation* plus global.

Les débats sur la place, la forme et la nature de l'observation dans une pratique empirique de la recherche, ne sont pas clos. Ils se prolongent au sein des Sciences humaines et sociales. La proximité, voire l'ambiguïté entre information et donnée avive ces débats dans une actualité dominée par les discours antagonistes et parfois schizophréniques sur les *Big Data*. L'échelle à laquelle l'observation se déroule n'est de toute évidence pas neutre. Les méthodes ou les outils d'analyses ont été conçus pour des volumes de données sans que l'on connaisse vraiment l'effet du changement d'échelle sur leur pertinence ou sur la qualité des résultats produits.

Le choix de l'observation expérimentale, associé aux dispositifs info-communicationnels et aux situations d'usage, ajoute une complexité qui n'est pas seulement liée à l'instrumentation et à sa nature technique. Cette complexité est inhérente au dispositif observé dont nous estimons qu'il ne peut pas être abordé dans un réductionnisme strict, séparant en deux entités distinctes un usager et un dispositif de médiation. Le sujet acteur est partie prenante du dispositif en fonctionnement que l'on observe. Il est par nature singulier et ne peut pas être réduit à un individu statistique. Pour l'observation de sujets, nous ferons l'hypothèse que chacun d'eux est une individualité distinguable parmi l'ensemble des entités agissantes du dispositif ainsi qu'une personnalité identifiée durant toute l'observation. Au contraire, l'environnement technique faisant dispositif est de moins en moins circonscrit et spécialisé. S'il est potentiellement permanent et étendu dans ces fonctionnalités (voire générique), ce sont les circonstances particulières de l'usage qui déterminent les configurations de son déploiement opérationnel.

Dans ce contexte, le terme d'observation est à comprendre dans un sens dépassant son emploi usuel, lié à la présence physique d'un observateur dans la situation observée. L'observateur peut en effet être éloigné dans l'espace et dans le temps (différé). L'intérêt (attendu) de la démarche que nous proposons est à la fois de pouvoir gommer la présence de l'observateur et d'habiliter l'observation directe continue pour des situations où l'observateur ne pourrait pas être présent. L'intérêt est aussi d'atteindre une acuité d'observation de la situation qu'un humain ne saurait ni égaler ni restituer.

L'instrumentation de l'observation ainsi que l'autonomie du dispositif observant n'est pas considéré d'un point de vue radical, comme une alternative à l'observateur mais comme une source d'observation complémentaire. Nos apports se situent dans une extension des domaines de l'observation et de l'observable. Il n'est donc pas question de redéfinir les paradigmes des épistémologies positivistes ou constructivistes mais de raisonner les extensions qui en sont possibles. Cette réflexion méthodologique et épistémologique accompagne notre travail et sera développée dans le chapitre suivant.

1. Constituer une unité informationnelle d'observation

La démarche expérimentale que nous poursuivons consiste à questionner ce qu'il est possible de capter de l'activité et des relations des acteurs humains ou non humains assurant collectivement le fonctionnement d'un dispositif. Nous ne reviendrons pas ici sur la *finalité* du dispositif ni sur *l'intentionnalité* sous-jacente constituant des hypothèses fortes de la logique d'usage. Nous prenons pour acquis que le fonctionnement du dispositif observé est pertinent et qu'il fait sens pour au moins l'un des acteurs agissant (ou usager). En particulier nous prenons comme postulat que les dispositifs apparaissent avec suffisamment de relief pour qu'ils fassent consensus (en première approximation) et qu'il n'y ait pas de doute sur les ressources qui les constituent[200].

La caractéristique de notre démarche est de n'envisager l'observation qu'au travers des flux d'informations associés au dispositif en fonctionnement. Bien évidemment, nous n'oublions pas que ce n'est pas le réel qui est observé mais des représentations complexes qui entretiennent avec le réel une proximité qui ne peut être mesurée. Par ailleurs, les représentations produites par le dispositif observé ainsi que les conditions de leur production ne sont pas toujours accessibles ni complètement documentées. Il y a donc une distorsion dans l'observation que nous ne pouvons pas négliger.

Dans les chapitres précédents, nous avons évoqué un principe d'enveloppement du dispositif étudié par le dispositif d'observation. Ce principe signifie que nous restreignons la mise en œuvre du dispositif d'observation aux éléments du dispositif observé que nous qualifions en conséquence de ressources[201].

À l'issue du chapitre précédent nous distinguons deux aspects de l'observation. En premier lieu, celui de *traçage* expérimental qui consiste à définir des mécanismes de sondes permettant de représenter des états de fonctionnement, et de capter des représentations informationnelles produites par les ressources actives du dispositif observé[202]. En second lieu, celui de *collecte* qui consiste à organiser des représentations en collections afin de favoriser l'exploitation des données qu'elles contiennent.

L'objectif de ce paragraphe est de définir l'*unité représentationnelle* enregistrée qui constitue la collection. Au sens informatique, il s'agit d'une structure de données décrivant l'unité d'enregistrement de l'information dans la collection. Cette unité est élaborée à partir de l'ensemble des traces numériques issues des sondes mises en œuvre durant l'observation. Elle correspond à une segmentation (discrétisation) et une sélection dans le flux d'information issu du dispositif observé. Elle est enfin porteuse des choix représentationnels caractérisant la finalité de la collection. Au regard des objectifs d'analyses, nous montrerons d'une part, que cette définition d'unité

---

[200] Nous mettons de côté le questionnement sur le "faire dispositif" *a prirori*. L'analyse issue de l'observation doit cependant permettre d'évaluer la consistance du dispositif.

[201] Une ressource peut être un dispositif.

[202] Rappelons qu'un dispositif est considéré comme un agencement dynamique de configurations d'entités actives (cf. chapitre 2).

d'enregistrement gagne à être enrichie d'informations contextuelles et d'autre part, que plusieurs niveaux d'organisation entrent en jeux dans la mise en œuvre d'une collection. Les choix représentationnels opérés au plus fin de la définition des collections font partie de la réflexion méthodologique qui accompagne la démarche d'observation.

1.1. Structurer l'information et les données

La structuration apportée à l'information dans la collection peut être considérée comme un compromis réalisé entre deux aspirations distinctes. D'un côté, celle de rendre compte de la structure du dispositif et de ses logiques de fonctionnement. De l'autre, celle d'anticiper le devenir de la collection et les analyses qui lui seront appliquées.

1.1.1 Structuration relative au dispositif observé

La structuration interne du dispositif observé est supposée dépendante des logiques d'organisation propres qui régissent son fonctionnement. Seule l'interprétation des manifestations informationnelles accessibles du dispositif nous permettent de tenter de reconstituer les logiques internes à ce dispositif.

Nous formulons néanmoins le principe que toute connaissance *a priori* disponible sur le dispositif et son organisation interne doit être exploitée au plus tôt, notamment dans les choix représentationnels visant la collection. Toutefois, ce principe ne pouvant pas être systématiquement mis en pratique, nous sommes conduit à partir du degré le plus faible de structuration, à savoir chacune des entités actives et productrices d'information à l'intérieur du dispositif. Cela revient à ignorer au moins au départ, les échelons intermédiaires en considérant qu'ils interviennent dans la mise en place itérative de la collection.

De manière générale, les entités humaines impliquées dans le dispositif sont individualisées et ne peuvent être décrites qu'au moyen de sondes placées en interface des entités du dispositif avec lesquelles elles interagissent. Les entités numériques[203] du dispositif sont assimilées, à des automates de traitements. Ils sont ainsi caractérisés par leurs états, leurs données d'entrées et de sorties disponibles ou susceptibles de l'être dans une interface. Chacun de ces types d'information fait l'objet d'une représentation numérique[204] horodatée (cf. §1.2).

Enfin, suivant la logique de contribution et de collaboration mise en œuvre entre les entités actives du dispositif, la production informationnelle globale du dispositif est aussi une entité représentationnelle signifiante et doit être conservée.

[203] Par hypothèse toute entité non humaine est de nature numérique ou peut faire l'objet d'une numérisation suffisante pour être considérée comme telle.

[204] Nous ne nous étendrons pas sur les caractéristiques interopérable et réutilisable des formalismes adoptés. Nous faisons implicitement référence au formalisme JSON largement déployé et répondant à ces objectifs.

1.1.2. Représentation adaptée à l'analyse

L'adaptation de la collection aux objectifs d'analyse consiste à produire des données conformes aux attendus des outils et des méthodes d'analyse envisagés. Ces attendus induisent une rétro-ingénierie qui oriente les choix de représentation. Les choix représentationnels s'inscrivent ainsi dans un double mouvement d'opportunités et de contraintes. C'est également en fonction de ces choix déterminants que la configuration du dispositif d'observation va être élaborée.

Il est difficile de raisonner ces contraintes en toute généralité et avec beaucoup de précision. Cependant, le premier des attendus analytiques est de produire des corpus[205], c'est-à-dire des ensembles de données significatives collectivement. S'agissant d'étudier des médiations info-communicationnelles[206], les corpus produits s'organisent autour de jeux de variables principalement associées aux sujets (personnes) mais aussi à d'autres caractéristiques du dispositif et de la situation observée, notamment ses dynamiques.

La première des contraintes est de pouvoir exprimer l'appartenance à un corpus d'un objet représentationnel (c'est-à-dire une entité numérique structurée) quel qu'il soit dans les termes de la représentation propre à la collection. L'objet représentationnel d'un corpus n'est pas toujours identique à un enregistrement de collection. Il peut être calculé à partir d'un ou plusieurs enregistrements. Ainsi, la représentation des données d'une collection doit être conçue pour permettre : l'identification d'enregistrement à partir de descriptions produites dans des termes de l'analyse ; les calculs nécessaires à la production des données de l'analyse. Ces conditions d'exploitation de la collection sont d'autant plus importantes que celle-ci est envisagée dans un contexte de massification des données. De ce fait, pour simplifier la génération de corpus et gagner en efficacité, certaines des données structurées du corpus peuvent être intégrées à la définition de la collection et calculées (si nécessaire) de manière anticipée.

1.2. Composition des représentations temporelles

La production de contenus par les ressources est généralement estampillée par des métadonnées temporelles permettant une lecture événementielle de la création et de la mise à disposition des contenus. L'horodatage des données est réalisé à partir des horloges des systèmes informatiques. L'horodatage instantané des représentations numériques est une pratique que l'on peut considérer comme systématique dans les ressources. Ainsi, l'ordre établi entre les contenus disponibles d'une même ressource est supposé être l'exact reflet de la chronologie de leur production. À l'échelle du dispositif nous assumerons l'hypothèse que les ressources sont alignées sur un même référentiel temporel universel.

---

[205] Rappelons que nous distinguons la collection qui décrit un objet documentaire associé au processus d'observation du corpus spécifique à un projet analytique le plus souvent fondé sur les données de l'observation.

[206] Soulignons que les collections peuvent répondre à d'autres projets d'analyse dont notamment le dispositif en tant que système complexe.

L'intégration temporelle uniformisée des différents événements construit une ligne de temps référente (*timeline*) à partir de laquelle l'ensemble des analyses temporelles est réalisée. La granularité de cette ligne de temps doit être la plus fine possible (généralement la seconde) pour supporter la plus grande variété de grilles (minute, heure, etc.) en fonction des objectifs analytiques et des phénomènes observés[207].

La difficulté à laquelle se heurte le *monitoring* vient de l'existence de latences possibles ou d'interférences dans la production des traces d'usage. Ces aléas, qui ne sont pas toujours prévisibles ni maîtrisables, peuvent se traduire par un schéma temporel aberrant du point de vue des logiques de l'activité.

Pour illustrer notre propos, considérons deux ressources A et B appartenant à un dispositif. Ces entités permettent la réalisation d'ensemble d'actions identifiées, respectivement : $\{a_1,\ldots,a_A\}$ et $\{b_1,\ldots,b_B\}$. Faisons l'hypothèse que le sujet produise une séquence d'actions correspondant à l'enchaînement de a_i suivi de b_j après un délai ε. Supposons que $E_x(t)$ formalise pour toute ressource du dispositif la durée d'effectuation de l'action x, débutée à l'instant t. Soit M_x la captation (mesure) effectuée à l'occasion du *monitoring* d'une entité du dispositif. Il faut alors que : $M_A(E_{a_i}(t)) < M_B(E_{b_j}(t+\varepsilon)+\varepsilon)$. Si ces conditions ne sont pas respectées, l'information d'accomplissement de b_j sera antérieure à celle de a_i et la compréhension que l'on aura sera erronée.

Dans la pratique, le délai ε qui dépend du sujet agissant, n'est pas toujours minorée par $E_{a_i}(t)$. Autrement dit, l'utilisateur n'attend pas forcément l'accomplissement total d'une action pour entreprendre la suivante. Ces anticipations d'actions se produisent lorsque le contexte d'exécution offre un feedback suffisamment explicite pour que l'utilisateur soit assuré de l'enchaînement effectif de a_i, puis de b_j.

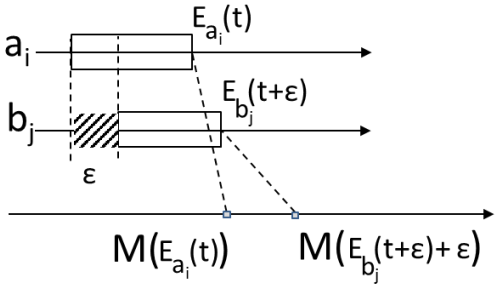

Fig37. Séquencement temporel des actions du sujet

Si l'anticipation est importante, le *monitoring* peut devenir aléatoire et ne plus restituer les séquences dans leur chronologie originelle.

Nous avons rencontré cette difficulté avec PLEXUS dans la mise en œuvre des jeux de cadres sur le Web du fait d'anticipations de navigation par les utilisateurs (Ref.16 p80). La structure informationnelle du navigateur nous a permis de repérer ces cas et de les redresser dans une phase

[207] Les grilles diffèrent suivant que l'on étudie, par exemple, un débat télévisé ou une campagne électorale.

de préparation des données. Mais cela n'est pas toujours possible. Le risque est de ne pas identifier les aberrations ce qui a comme conséquence de biaiser l'interprétation.

La réponse technique consiste à mettre en œuvre une sonde plus sophistiquée, permettant une granularité de restitution plus fine que la simple chronologie d'actions. Mais cela ne peut s'envisager que si l'interaction avec les entités du dispositif est réglée par un protocole de communication ayant lui-même une granularité adaptée permettant de signaler la prise en charge et la finalisation de l'action au moment de son effectuation.

Dans ces conditions, on peut reformuler l'effectuation sous la forme :

$E_x(t) = \Delta_x + C_x(t) + \Phi_x$. Dans cette formule, $\Delta_x(t)$ traduit le délai de prise en compte de la demande de x ; $C_x(t)$ correspond à la charge effective qu'engendre le traitement de l'action x à l'instant t ; et $\Phi_x(t)$ le délai d'accomplissement finalisant x. Ce cas de figure correspond à un protocole de communication du type HTTP. Celui-ci se traduit par un balisage temporel fin signalant les deux étapes de prise en compte et d'accomplissement[208]. Dans le cas d'une page Web, par exemple, Δ_x correspond à la sollicitation de la page et Φ_x correspond au délai d'affichage sur le moniteur.

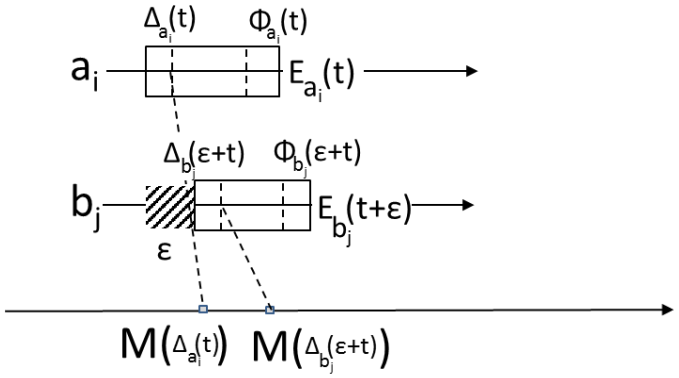

Fig38. Balisage temporel des événements de communication

L'anticipation étant le plus souvent fondée sur un signal associé à $\Delta_{a_i}(t)$, il y a de ce fait, peu de risques d'inversion : en rapprochant le point d'ancrage de l'activité de son origine, on gagne en précision. Le protocole n'évite cependant pas tous les aléas de fonctionnement mais les contrôles introduits réduisent considérablement les risques d'erreur. En outre, l'amélioration apportée par cette décomposition se traduit par une plus grande capacité d'analyse de l'activité dont une appréciation de la durée devient possible.

D'un point de vue général, ces considérations renvoient à la détermination d'une granularité informationnelle et temporelle adaptée aux logiques d'interaction entre les entités. Cela signifie que les interactions sont considérées et décrites en tant que phénomènes de communication. Les problèmes soulevés proviennent de la présence d'acteurs humains qui bien qu'agissant séquentiellement sont susceptibles d'anticipation. Ces contraintes doivent être intégrées dans la

[208] Voir (Ref.16, p81).

mise en œuvre d'outils expérimentaux portant sur l'étude des interactions instrumentales suivant différents canaux indépendants dès lors que la synchronisation n'est pas contrôlée par une interface.

Le contexte est différent pour les configurations de dispositifs que nous avons étudiées avec MEDIASWELL. En effet, les sujets n'ont été observés qu'indirectement, au travers des représentations que les ressources élaborent elles-mêmes lors d'interactions individuelles. Ces représentations ne documentent pas ce qui se rapporte à l'instrument mais ce qui se rapporte à sa fonction médiatrice. Les effets d'anticipation, notamment dans l'interaction interpersonnelle (dialogique) sont neutralisés. Dès lors, la ligne temporelle devient la base de la synchronisation des différents événements constitutifs de l'activité globale au sein du dispositif.

1.3. Contextualiser l'unité d'observation

La prise en compte de la situation d'usage suppose de ne pas replier l'observation uniquement sur le dispositif mais d'identifier ce qui en constitue l'écosystème et plus globalement son environnement. La notion de dispositif incite à une approche systémique. Considérer ce dispositif comme un *système ouvert* est clairement apparu comme nécessaire à l'occasion des expérimentations portant sur des événements médiatiques. L'utilisation de références très nombreuses à des ressources audio-visuelles issues d'Instagram, Facebook, FlickR, Youtube, etc. dans le corps des messages (Tweets) publiés dans Twitter peut être considérée soit comme l'appartenance de ces ressources au dispositif médiatique étudié soit comme l'interaction du dispositif avec son écosystème. Dans certains cas, les relations contractuelles entre les acteurs de ces services permettent un arbitrage en faveur du premier cas. À défaut d'un point de vue fixé par le projet d'étude, ce sont les logiques d'usage qui nous semblent devoir guider ces arbitrages dont il faut conserver la mémoire.

La prise en compte de l'écosystème dans la collection s'est imposée du fait de la volatilité des informations dans l'Internet, leur complétude ne pouvant être assurée qu'au moment de la collecte. À la suite de nos travaux, notamment chez PRÉDICTYS[209], nous avons généralisé cet enrichissement en l'intégrant dans un principe de contextualisation de collection. L'écosystème constitue un périmètre contextuel naturel puisque le dispositif s'y enracine. Sa détermination est en général relativement intuitive. À des fins heuristiques, il peut être pertinent de repousser ce périmètre pour inclure des ressources qui ne sont, *a priori,* pas explicitement reliées au dispositif ou à son écosystème. Il s'agit de ressources dont l'observation simultanée *semble* intuitivement pertinente.

La collecte de ressources associées à l'écosystème s'envisage de deux manières.

La première est guidée par les données collectées. Elle consiste, à identifier dans ces données celles qui font référence à d'autres ressources de l'écosystème, et à étendre la collecte aux résultats des requêtes produites à partir de ces données auprès de ces ressources. Le cas typique est celui de la

[209] Nous avions constaté que le système des données de production ne permettait pas à lui seul de raisonner les données. Il est par nature endogène et n'apporte que peu d'information sur la ressource. Nous nous sommes ainsi convaincu de l'intérêt d'ouvrir le système pour y incorporer des données exogènes.

récupération des fichiers d'images (*bitmap*) sur Instagram à partir d'une URL *tweetée*. Il s'agit comme on le constate d'une extension à la demande de la représentation.

La seconde est guidée par le *monitoring* systématique de l'intégralité des ressources de l'écosystème. Cette méthode se justifie si la compréhension des logiques internes à chacune des ressources est jugée indispensable dans l'analyse. Cette manière de procéder doit cependant être très encadrée. En effet, suivant la ressource, celle-ci peut être extrêmement productive ; ce qui peut dénaturer la collection (déséquilibre de représentation) ou dégrader les performances de la collecte.

Le choix de l'une ou l'autre des deux méthodes relève le plus souvent d'une économie de projet (durée) et de ressources (volume). La première est plus économique que la seconde, mais cette dernière est potentiellement plus féconde. L'enregistrement des flux de l'AFP réalisé à l'occasion des Jeux Olympiques de RIO entre dans cette logique. L'AFP est une ressource de l'écosystème médiatique dont l'importance n'a plus à être démontrée. Pour cette ressource, nous n'avons pas enregistré l'intégralité des publications produites (API) mais seulement celles qui concernaient l'événement olympique, c'est-à-dire le thème de la collection.

Les informations ainsi extraites de l'écosystème viennent enrichir les représentations issues du dispositif en fonctionnement. Ces apports doivent être distingués (structures de données ou métadonnées) afin de conserver une lisibilité des périmètres respectifs. La distinction apportée ne doit cependant pas être préjudiciable à l'unité des objets représentationnels que l'écosystème complète ou articule.

2. Produire des collections enrichies pour des corpus documentées

De par sa conception modulaire, MEDIASWELL intègre une grande diversité de ressources, ce qui introduit une généralisation du concept de trace numérique d'usage nous amenant à mieux articuler les notions de données et de métadonnées de collection. Cette ouverture sur des ressources diverses assure une prise en charge de projets d'analyse variés. Ainsi est-il plus facile pour nous de proposer des collaborations et de nous inscrire dans les dynamiques de la recherche en Sciences humaines et sociales. L'extension du domaine d'application, comme celui de l'interdisciplinarité vont de pair avec une exigence accrue de communication, de mise en circulation des données au sein du projet et en dehors. Ce mouvement d'extériorisation des données de la recherche les détache du contexte expérimental, du dispositif de collecte et de sa configuration. Cette perte d'ancrage nécessaire doit être prise en charge au niveau des corpus de données afin d'en garantir l'unité et l'autonomie. L'enjeu d'indépendance des corpus de données produits s'exprime ici en tant que qualité d'un système de collecte. Cet enjeu pratique trouve un écho plus global dans les discours sur la publicisation des données que portent les institutions de la recherche.

2.1. S'inscrire dans une dynamique scientifique

2.1.1. Anticiper l'évolution du projet analytique

L'observation expérimentale intervient souvent dans une étape exploratoire d'une recherche empirique. Dans ce contexte, les attentes associées à une collection de données sont le plus souvent très générales. Le mot d'ordre est, "observons et nous verrons bien…". L'observation est investie d'un pouvoir de révélation, les données semblent devoir surgir du terrain et en assurer une lisibilité quasi immédiate. Malheureusement, l'enchantement n'est jamais au rendez-vous, et la magie des données inopérantes. Ainsi faut-il souvent amorcer une collection puis par raffinements successifs, évaluer et reconsidérer l'ensemble des ressources explorées ainsi que la nature des informations collectées.

La collecte et l'évaluation nécessitent une durée et une disponibilité importante des observateurs/analystes pour être menée à bien. Dans l'économie générale du projet, ce temps est jugé faiblement productif puisqu'il se situe en amont de l'analyse, elle-même considérée comme le cœur du projet. De plus, le caractère souvent immédiat ou éphémère des phénomènes que l'on veut observer (cas des événements médiatiques) ne permet le plus souvent que des ajustements à la marge. Enfin, le projet d'analyse peut lui-même s'écarter des objectifs initiaux ou engendrer des prolongements mobilisant des données non prévues. Aussi, il ressort de ces différents éléments la nécessité d'anticiper les besoins de données et de collecter large. Ainsi l'adaptation aux besoins analytiques relève d'un second temps qui est celui de la constitution du corpus de données.

À titre d'exemple, dans le projet d'analyse des trajectoires individuelles de professionnels (journalistes) conduit à partir des pages de profils LinkedIn, la plupart des informations de la page publique ne sont pas utilisées. Cependant, conserver une représentation de la page complète nous a permis d'émettre des hypothèses *a posteriori* concernant les compétences associées aux profils.

2.1.2. Associer des métadonnées expérimentales

Selon l'Association des Documentalistes et Bibliothécaires Spécialisés (ADBS), les métadonnées sont définies en tant qu' « *ensemble structuré des données créées pour fournir des informations sur des ressources électroniques »*[210]. Ce sont des données de nature descriptive répondant à un principe documentaire et dont la signification est associée à une logique externe d'utilisation. Le fait de ne considérer que des *ressources électroniques* est une restriction historique associée à l'emploi du terme métadonnée dans la mise en place du Web. Si ce terme peut être généralisé à toutes sortes d'entités, la définition que l'on trouve parfois de « *données sur des données »* est incomplète et à proscrire[211].

Dans le cas de la définition d'un dispositif d'observation expérimental, les métadonnées ont comme fonction de rendre compte du processus d'observation dans sa durée. Dans le cas présent, l'observation est associée à la collecte de traces numériques d'usage produites par ou à partir de ressources d'Internet. Nous distinguons ainsi quatre types de métadonnées :

---

[210] http://www.adbs.fr/metadonnees-17808.htm?RH=OUTILS_VOC

[211] Elle conduit à une *datafication* ou mise en données du monde qui ne nous apparaît pas satisfaisante intellectuellement.

1. Celles qui caractérisent les ressources observées. Ce sont des métadonnées descriptives portant à la fois sur des attributs stables dans la durée ou variables suivant le contexte. Ces métadonnées portent sur leurs caractéristiques spécifiques des ressources, sur les modalités de leur utilisation ainsi que sur leur production. Compte tenu de la durée importante de l'observation, ces métadonnées descriptives sont nécessairement horodatées et associées à des instants particuliers de la collecte ;
2. Celles qui décrivent le contexte expérimental. Il s'agit d'une description documentaire externe venant caractériser l'expérience elle-même ainsi que son inscription dans le projet global d'analyse de données.
3. Celles qui définissent le dispositif observateur dans son fonctionnement. Ces métadonnées décrivent la configuration de modules fonctionnels mis en œuvre et leur activité. Ces métadonnées sont expressément définies pour rendre compte du processus d'observation le plus finement possible. Elles assurent une double finalité de *monitoring* du processus de collecte et de documentation de la collection. Elles ont pour objectif de rapporter les conditions particulières de fonctionnement et les aléas de la collecte de données. Les valeurs associées sont produites par les modules fonctionnels actifs, à des instants clefs des traitements qu'ils effectuent ;
4. Celles qui concernent les enregistrements de collection et les valeurs déterminées aux différentes étapes de la collecte de données ou de leur complétion. Il s'agit de métadonnées qualitatives dont l'objectif est de rendre compte de la confiance (indice) que l'on peut accorder à l'information associée à la valeur collectée ou calculée. L'exemple de la détermination du sexe à partir du prénom illustre de notre point de vue cette nécessité[212].

Nous avons fait le choix d'archiver les métadonnées dans la structure de données des enregistrements de collection. Nous n'avons pas normalisé la déclaration des métadonnées que ce soit dans le choix d'une nomenclature d'attributs, ou dans le choix de structures de données. À notre connaissance, les recommandations proposées par le World Wide Web Consortium (W3C) ne concernent encore actuellement que de bonnes pratiques[213]. Cependant, la publicisation des corpus impose la mise en œuvre d'une norme partagée ou d'une documentation suffisante des choix représentationnels.

2.1.3. Publiciser les données de la recherche

La publicisation des données de la recherche apparaît désormais comme un corollaire de plus en plus indispensable de l'acte de publication. Au début des années 2000 nous avions été particulièrement attentifs à la mise en place des archives ouvertes (Ref.10, 14, 15). La mise à disposition des corpus prolonge cette initiative scientifique. Les objectifs sont similaires et relèvent à la fois, de la méthodologie, de l'économie et de l'éthique scientifique. Par exemple, l'administration de la preuve, renforcée par la disponibilité des données est à la fois d'ordre méthodologique et

[212] L'utilisation de la base INSEE des prénoms permet de mettre en place une métrique de ce type. Nous l'avons expérimentée à de nombreuses reprises dans nos travaux.

[213] https://www.w3.org/TR/dwbp/ version du document 19 mai 2016 consultée le 12 juillet 2016.

éthique. De même, la subsidiarité des productions scientifiques, assurées par la mise à disposition des données et des résultats de la recherche correspond à des logiques méthodologique et économique. Nous sommes ici dans une lecture d'ordre scientifique. Cependant, la publicisation des données va aussi dans le sens d'un projet politique global de contrôle et de transparence de la dépense publique en matière de recherche et de son efficacité scientifique. Les mouvements de l'*open data* ou de l'*open access*, ont donc une double légitimité scientifique et politique.

Bien que des directives européennes aillent clairement vers l'organisation de la mise à disposition des données de la recherche, notamment dans le cadre des programmes de financement de la communauté européenne (H2020)[214], ce projet n'est toujours pas abouti. Cela ne nous empêche pas d'intégrer ces logiques.

Dans ce contexte la définition d'un plan de gestion des données (PGD) ou *Data Management Plan* (DMP) est une étape clef de la publicisation. Un DMP[215] peut être considéré comme un document administratif de gestion de projet. Le DMP est un document d'accompagnement qui est susceptible d'évoluer au fur et à mesure de l'avancée de la recherche. À ce titre, il a comme objectif d'isoler, dans la conduite du projet, les éléments relatifs à la production et à la gestion des données de la recherche. Cette manière de faire exister les données dans le projet de recherche permet de leur attribuer une valeur et d'entrer en quelque sorte dans une généralisation de l'économie des données.

La publicisation des données fait partie des objectifs clairement associés au DMP. Pour les corpus rendus publics, le DMP doit comporter les différentes informations administratives, juridiques, documentaires et techniques nécessaires à l'exploitation ultérieure des données.

La démarche formelle associée à la publicisation des données peut aller très loin. Nous nous restreignons à l'usage de métadonnées descriptives[216], et en premier lieu, aux données décrivant les licences, la provenance, la qualité, la version.

Si ces informations ne se traduisent pas toutes par des contraintes lors de la production de corpus, certaines imposent des conditions particulières dans la mise en œuvre des collections. Par exemple, le choix d'un formalisme de mise à disposition peut être adopté lors de la production du corpus. En revanche, les conditions légales correspondant aux données personnelles (anonymisation, etc.) sont des contraintes fortes, et doivent être prises en charge dès la collection[217].

D'une façon générale, et indépendamment de l'aspect contraignant, il est préférable d'inscrire la plus grande partie des informations du DMP dès la constitution de la collection. En effet gérer les informations du DMP au plus près des données de collecte, nous semble être non seulement une

---

[214] http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html. La mise en place d'un projet pilote : *Open Research Data* souligne cet intérêt et anticipe ce que sera probablement une bonne pratique dans la recherche financée sur des fonds publics d'ici quelques années.

[215] De nombreuses ressources deviennent disponibles : https://dmptool.org/

[216] On peut en effet documenter les structures de données utilisées pour les enregistrements, les corpus ou les collections, ce qui devient particulièrement lourd à mettre en œuvre.

[217] Tout du moins dans un délai court ce qui incite à structurer les données de collection pour faciliter cette contrainte contrôlée par la CNIL.

bonne pratique mais également un moyen d'organiser le cycle de vie des collections. De ce point de vue, le DMP n'est pas qu'un artifice administratif, une injonction normative, institutionnelle et tutélaire. Il a du sens dans une perspective de recherche. Autrement dit, documenter le processus de collection est une étape naturelle relevant de la méthodologie expérimentale.

Nous faisons usage du terme DMP[218] dans le contexte restreint de la publicisation des données. Nous l'utilisons pour désigner des métadonnées descriptives structurées, à la manière de celles qui interviennent dans les protocoles de moissonnage (Ref.14). Dans la pratique, les DMP traduisent d'une part, les enjeux généraux de communication et de valorisation associés à un projet de recherche et, d'autre part, les enjeux méthodologiques et expérimentaux liés à des projets d'analyse. Seuls les derniers retiennent notre attention car ils relèvent de la définition du dispositif et des modalités de collecte.

Pour aller plus loin dans la mise en commun des données, au-delà des enjeux d'interopérabilité et d'intelligibilité, il s'agit de mettre en place les moyens d'une exploitation étendue favorisant les collaborations et la réutilisation des collections. Il convient de dépasser le strict respect des normes et d'enrichir les données d'informations sémantiques.

Nous suivons en ce sens les propositions de Tim Berner Lee, dont les travaux se rapportant aux données liées (*linked data*) font l'objet d'un important travail de sensibilisation et de développement au sein du W3C.

La mise en place de relations entre les données (voire métadonnées) suppose l'existence d'un modèle interprétatif associé à la collection supportant une lecture la plus générique possible en dehors du cadre d'analyse. Cela suppose en particulier d'avoir recours à des ontologies du Web stables et faisant autorité. Nos travaux nous ont ainsi conduit à explorer la piste de DBPédia afin d'identifier les entités nommées. Pour l'établissement d'une *sémantisation* plus poussée des collections, le chantier reste ouvert.

2.2. Enjeux méthodologiques de la production de données

Dans ce paragraphe, nous définissons par méthodologie ce qui renvoie aux questions de moyens et de nécessité, c'est-à-dire ce qu'il convient de mettre en œuvre pour produire des données de collection à partir d'un dispositif à observer. Ce versant méthodologique[219] fait référence à des méthodes d'élaboration et d'instanciation de structures de données à partir de traces d'usage numériques publicisées ou extraites de ressources info-communicationnelles. Ce questionnement a été abordé dans les chapitres portant sur les traces d'usage (chapitre 3) et les dispositifs de traçage (chapitre 5), ainsi que dans le début de ce chapitre. Il s'agit désormais d'en établir la synthèse.

Cette réflexion repose sur l'hypothèse d'existence d'un espace numérique global très étroitement associé à l'espace d'activité humaine par un ensemble de dispositifs info-communicationnels dans lesquels des personnes sont impliquées. La réduction problématique opérée en étudiant

[218] La définition des DMP est encore peu fixée.

[219] L'autre, concernant la production de connaissances sera abordé au paragraphe suivant (§3.3).

uniquement Internet, sous-espace normatif de cet espace numérique global, est ici sans conséquence du point de vue méthodologique. Les logiques d'usage sont neutralisées à ce niveau.

Comme nous l'avons souligné au paragraphe (§1) la dimension méthodologique de l'observation se décompose en deux phases. D'une part, le *monitoring* du dispositif de médiation observé et d'autre part, la constitution d'une unité de collection (§2).

2.2.1. Monitorer et produire des données

La disponibilité temporelle des ressources, leur dynamique ainsi que leur régime de production informationnelle affectent la collecte de données qui doit être extrêmement réactive. La déclinaison en processus autonomes et parallélisables que nous avons systématisée dans MEDIASWELL est une réponse adaptée. L'utilisation de mécanismes de files d'attentes vient compléter cette proposition méthodologique dont l'objectif est de coller aux dynamiques du dispositif afin de ne pas altérer le flux d'information et d'en réaliser une capture fluide, sans perte, la moins différée possible (selon la capacité d'absorption). Ces préconisations techniques suggèrent la mise en œuvre d'architectures fortement modulaires et distribuées qui sont productrices à leur tour de flux d'informations de contrôle également monitorés. Il s'agit en définitive de produire un dispositif expérimental qui remplisse les conditions de sa propre observation.

La constitution d'une unité de collection consiste à intégrer et organiser l'ensemble des traces numériques produites par les ressources actives ou les sondes du dispositif observé. Dans le paragraphe 1.2 nous avons insisté sur les difficultés liées à la synchronisation événementielle dans des environnements distribués. La composition des représentations doit supporter les alternatives que l'incertitude temporelle engendre. Cette caractéristique va dans le sens d'une généralisation de la traçabilité des ressources (origine) constitutives de l'unité de collection (enregistrement).

La dimension chronologique, sérielle ou temporelle, est très présente dans le principe de collection. Le *monitoring* coordonné mis en œuvre dans le dispositif d'observation assure la synchronisation et l'uniformisation des références chronologiques dans la représentation. Il est ainsi possible d'envisager la collection comme un processus incrémental que nous exprimons dans la série suivante où $C()$ désigne la collection réalisée, et $U()$ l'unité informationnelle collectée à l'instant t_n [220]: $C(n) = C(n-1) + U(t_n)$ où $C(0) = U(t_0)$.

L'opérateur "+" traduit l'actualisation de la collection à l'issue du nième pas de collecte. Nous pouvons formaliser l'unité de collecte par les éléments représentationnels qui le constituent. Chacun de ces éléments comporte des métadonnées descriptives.

Nous distinguons ainsi deux sortes de représentations enrichies :

- $R_i(t)$: associée à une ressource i du dispositif d'observation.

[220] L'instant est un horaire lorsque la capture est passive : l'information est alors poussée par la ressource qui produit des données. Il est un indice dans une série lorsque la capture est active, inscrite dans un cycle contrôlé.

- $C_{j,k}(t)$: associée à une ressource contextuelle j du dispositif d'observation. k=0,1 en fonction de la nature générique (0) ou écosystémique (1).

Nous noterons $U(t) = \bigcup_i \bigcup_{j,k} R_i(t) C_{j,k}(t)$

2.2.2. Technique et compétences

La mise en œuvre de sondes relève d'une démarche expérimentale d'observation requérant une connaissance événementielle très fine des processus internes du dispositif de médiation. Il s'agit d'établir les éléments signifiants parmi ceux susceptibles de survenir et d'élaborer les sondes ou de manipuler les boites à outils permettant de les capter. L'effort d'appropriation technique demandé est important et le plus souvent très spécialisé (en d'autres termes : faiblement réutilisable). Il n'est par ailleurs pas toujours possible d'entrer aussi profondément dans l'ingénierie du dispositif et l'on doit se contenter alors des événements externes produits par les ressources actives. Ce constat que nous vérifions dans nos travaux pose la question des compétences techniques disciplinaires et des collaborations complémentaires qui selon le cas s'imposent.

Les éléments méthodologiques que nous venons d'évoquer rejoignent ceux qui concernent l'étude des interactions entre acteurs et en particulier de type homme-machine.

La réponse souvent mise en avant dans les discours et parfois dans les travaux consiste à déléguer les aspects techniques soit dans une relation de sous-traitance (ingénieur) soit dans une collaboration interdisciplinaire. Ni l'une ni l'autre n'est une réponse facile. Dans les disciplines SHS où il existe une proximité forte avec le numérique, la collaboration ne peut s'établir efficacement qu'à partir du moment où le bagage technique commun est suffisamment riche de part et d'autre. Cet effort d'acculturation pluridisciplinaire et technique nous semple indispensable pour le devenir de ces disciplines des SHS qui risquent sinon d'être dépossédées de leur objet. Une alternative, tout aussi délicate consiste à collaborer avec les promoteurs des ressources afin qu'ils fournissent et autorisent l'exploitation des données qui leurs sont propres. Nous ne coupons pas à cette exigence qui tient à l'exploitation de données éventuellement publicisées mais dont le statut public n'est pas avéré. Ce sont les modalités et les clauses de la collaboration qu'il faut pouvoir rendre favorable à la rigueur et à l'intérêt des pratiques scientifiques.

3. S'inscrire dans une démarche empirique renouvelée

Le questionnement méthodologique que nous poursuivons dans ce chapitre, ne peut pas faire l'économie d'un basculement de point de vue qui nous conduit à aborder les collections dans le contexte d'une démarche empirique de production de connaissances.

Pour ce qui précède et qui concerne le cadre expérimental, l'ensemble de notre réflexion sur les aspects méthodologiques se nourrit des incessants allers-retours entre les deux activités de collecte et de fouille de données qui fondent notre pratique. Il s'agit de raisonner les éléments méthodologiques du point de vue de la production de données uniquement. La confrontation aux contraintes et aux enjeux de traitements est alors indispensable pour, en retour, affiner les choix de représentation et de structuration des collections. Ces apports, issus de l'expérience acquise dans le

traitement, traduisent une optimisation itérative. L'optimisation s'avère relative. La prise en compte des collections en tant que ressources publicisées (§2) met en évidence qu'elles résultent davantage d'un compromis entre computation et communication.

Dans cette nouvelle partie, le renversement de perspective que nous proposons de suivre amène à considérer la collection comme un tout. De ce point de vue, l'exploration des traces numériques d'usage ne présente d'intérêt à nos yeux que dans la mesure où l'analyse de collections constituées permet de changer d'échelle en passant de l'échelon *micro* aux échelons supérieurs. L'ambition est clairement d'atteindre l'échelon *macro*, c'est-à-dire de travailler avec des méthodes et des collections suffisamment bien dimensionnées et validées pour que les résultats produits atteignent la portée la plus générale possible. Adopter cet objectif nous permet de fixer un horizon sans pour autant perdre de vue la nécessité qu'il y a de consolider les bases méthodologiques de la démarche.

Au travers de cet objectif, nous nous situons au cœur d'un champ de questionnements que certains associent à un « *second virage computationnel* »[221] (Marres, Weltevrede, 2012, p2) voire à une « *troisième génération* » (Boullier, 2015) des Sciences humaines et sociales. Il s'agit d'envisager de quelle manière produire une réflexion scientifique à partir de données numériques dont les caractéristiques qualitatives et quantitatives ne répondent plus aux critères d'exhaustivité, de représentativité, de vérité, etc. établis antérieurement comme règles intangibles.

Les débats actuels portant sur l'exploitation massive des données numériques, s'articulent autour de la notion contestable mais néanmoins largement vulgarisée de *Big Data*. Cette expression traduit un dépassement multi dimensionnel de grandeurs associées à des collections de données. L'évocation précédente de l'échelon *macro* nous situe dans cette perspective. Il s'agit de questionner aux limites notre approche des collections et des corpus[222] à la lumière de la notion *Big Data*. Nous aurons l'occasion de revenir dans le dernier chapitre sur les débats épistémologiques qui lui sont associés.

3.1. Qu'entend-t-on par *Big Data* ?

L'estimation annuelle de la quantité d'informations numériques produites et circulant sur Internet se situe en 2015 au-delà du *zettabyte*[223]. Ce volume considérable, progressant de manière exponentielle, sans équivalent dans l'histoire de l'humanité incite à penser autrement le rapport à l'information et au document. Les technologies de l'information et de la communication à l'origine de cette explosion informationnelle uniformisent le point de vue, résumant la diversité et la nature de contenus produits et échangés en une unité numérique : la *donnée*. Ce terme est ici à comprendre

[221] *The second computational turn*

[222] Par la suite nous évoquerons de manière équivalente collection et corpus tout en sachant que la dimension computationnelle est essentiellement visée dans les corpus extraits et moins dans la collection dont nous privilégions le sens documentaire.

[223] https://fr.scribd.com/doc/58589040/Internet-in-2015 infographie CISCO. Consultée le 20 juillet 2016.

dans un contexte computationnel équivalent à celui d'une machine de Turing[224]. Il porte en lui le fantasme contemporain d'une équivalence entre information et donnée conduisant à une calculabilité généralisée des activités. Le terme *Big Data* apparaît au tournant des années 2000 et connaît une très grande actualité depuis les années 2010. On trouve également réparties, des références qui font DU *Big Data* un champ d'étude et de compétences ou DES *Big Data* une objectivation quantifiable de l'effet de l'activité humaine et sociale dans l'espace numérique[225]. Ces différentes expressions investissent l'ensemble des discours économiques, scientifiques et techniques. Elles illustrent une tendance croissante à la mise en données ou *datafication* des réalités du monde mais aussi un rapport nouveau à la société.

«… *l'ère du Big Data n'est pas une simple mode qui passera, elle est portée par des dispositifs techniques, institutionnels, cognitifs, marchands et des discours qui font système, qui font "matérialité" et "énoncé" comme tout dispositif (Foucault, 1966) pour produire une nouvelle offre d'interprétation du social »* (Boullier, 2015, p2).

3.1.1. Le volume comme légitimation et évidence

Le *Big Data* est souvent intuitivement associé à la possibilité de numériser, calculer et prévoir le monde. La question de *volume* est très présente dans les discours qui l'accompagnent. Dans les lignes précédentes, nous avons d'ailleurs considéré quantité et volume comme deux grandeurs isomorphes dans le contexte de l'information numérique. Cet implicite est fondé sur l'équivalence entre information numérique et inscription binaire. Le volume exprimant alors[226] le nombre de bits nécessaires à l'encodage.

Mais comme le souligne Lev Manovich ce que désigne l'expression *Big Data* est relatif à une époque donnée (Manovich, 2011). Ainsi, le concept de *Big Data*, à l'instar de celui de *nouvelles technologies* en son temps s'actualise en permanence du fait de l'augmentation constante du volume d'informations numériques et de l'amélioration continue des performances des dispositifs numériques de stockage, de communication et de traitement. Ainsi, la référence aux performances limitées d'un poste de travail n'est donc pas valide en tant qu'expression d'un seuil au-delà duquel débuterait le *Big Data*. L. Manovich ajoute également que les considérations portant sur le *Big Data* diffèrent selon que l'on situe les enjeux dans les domaines de l'industrie, des humanités numériques ou des sciences classiques (*Ibid.*, p2).

Pour les sciences de l'ingénieur, le calcul haute performance ou *High performance Computing* (HPC) constitue un défi en soi. Il est une clef permettant de modéliser ou de simuler des phénomènes naturels plus rapidement à partir de volume de données toujours plus grand. Dans le discours institutionnel du CNRS que l'on trouve reproduit dans des supports de présentation, comme dans le livre blanc consacré au calcul intensif, le *Big Data* « *qui est, depuis peu, considéré comme le 4e pilier de la*

---

[224] Nous associons le terme de "donnée" à la traduction du terme "*data*".

[225] Ce chapitre nous fait privilégier LE *Big Data* dont nous questionnons les méthodes et les techniques. Nous aborderons LES *Big Data* dans le chapitre suivant.

[226] Abus de langage probablement lié à la forme substantive de l'information.

science moderne (le troisième étant constitué par la modélisation et la simulation) » est décrit comme « *un enjeu majeur* »[227]. Dans ce contexte, le *Big Data* représente un horizon technique et technologique aux limites des capacités et de la performance des artefacts computationnels. Tout l'enjeu consiste à élargir cet horizon. La relativité du discours est en l'occurrence nécessaire pour actualiser en permanence ce défi.

L'importance scientifique du *Big Data* est alors une évidence justifiant les investissements. Enfin, le critère de volumétrie est suffisamment représentatif de classes de problèmes scientifiquement pertinents pour ne pas appeler plus de justifications. Cette orientation politique du discours sur le *Big Data* produit un effet d'alignement pour l'ensemble des communautés scientifiques y compris celles issues des Sciences humaines et sociales. Ainsi, les problématiques spécifiques jusqu'alors définies comme *humanités numériques* en subissent le contre coup et courent le risque de se perdre/diluer dans l'injonction de massification.

Un raisonnement assez analogue à celui du HPC est tenu par les acteurs dominant de l'Internet qui engendrent un trafic très important et qui possèdent de très gros volumes de données. Ces acteurs se sont dotés de moyens et de compétences (computationnelles et analytiques) par des jeux d'alliances, de rachats ou de recherches leur permettant de développer une activité aux frontières du *Big Data*. Dans leur cas, le *Big Data* est une opportunité de croissance et le caractère massif un avantage concurrentiel acquis.

De manière générale, dans le secteur économique, envisager les données sous l'angle du volume (massification) renvoie à la question des collections et de la valeur financière voire symbolique qu'elles représentent. Selon cette approche, le *Big Data* est révélateur d'enjeux industriels et marchands qui s'expriment en amont de leur exploitation, dans la constitution des collections. De ce point de vue, le *Big Data* est étroitement lié aux *systèmes d'information* qui organisent les données. Le *Big Data* est alors un objectif associé à l'efficience des systèmes d'information produisant en volume et en qualité des données utiles au fonctionnement de l'entreprise. Cette problématique stratégique de l'information-donnée est déjà ancienne et remonte au milieu des années 1980 (Porter, Millar, 1985). Dès cette époque, les enjeux liés à la maîtrise de l'information (interne et externe) ainsi que l'incidence de l'information dans la chaîne de la valeur sont clairement identifiés.

Pour autant, dans l'approche historique des systèmes d'information, le volume n'est pas considéré comme un avantage mais comme une contrainte qu'il faut gérer notamment en développant des stratégies sélectives de conservation et d'effacement des données. L'actualisation qu'introduit le *Big Data* met un accent positif sur la croissance en volume des données. Elle est principalement associée à l'émergence du Web en tant qu'espace globalisé de production et de consommation d'information et de services. C'est dans cet écosystème numérique et globalisant que l'économie de la donnée prend désormais sens.

---

[227] Préface d'Alain Fuchs président du CNRS. Livre blanc http://www.cnrs.fr/ins2i/IMG/pdf/Livre_blanc_-_derniere_version.pdf.

Avec l'approche marchande, deux points de vue coexistent désormais. Le premier est associé à la production et se traduit par un déplacement des logiques de l'industrialisation des biens de consommation dans l'espace numérique. Il s'agit de produire des services d'information, valorisant un savoir-faire technique sur les données ou de monétiser des données valorisées (enrichies, attestées, etc.)[228]. Le second est de nature stratégique et voit dans le Web une extension de l'espace d'information de l'entreprise. Dans les deux cas, le volume est considéré comme un élément d'appréciation de la valeur, celle-ci augmentant avec la taille de la collection.

La réflexion en volume de la valeur informationnelle n'est pas propre au monde de l'entreprise ; elle affecte l'ensemble des espaces productifs. Produire plus ou recueillir davantage d'information devient gage de performance. Cette appréciation quantitative de la performance entretient un rapport ambigu avec la pertinence du résultat. L'ambiguïté repose en partie sur des représentations hypostasiées de la computation de l'information, du calcul numérique et des statistiques. Elle est entretenue par les producteurs du marché des données qui trouvent dans la quantité un argument commercial mais aussi une mesure adaptée aux enjeux de la monétisation. Elle est supportée par l'ensemble des acteurs économiques et politiques qui trouvent dans la mesure une réponse au problème de l'évaluation. La quantité est enfin une caractéristique liée à la conduite d'actions commerciales qu'à défaut d'un ciblage individualisé on ne sait véritablement mener aussi bien efficacement que massivement (Ref.17).

3.1.2. Au-delà du volume

Le critère de volume ne suffit pas pour rendre compte des enjeux associés au *Big Data*. La littérature sur la question fait apparaître d'autres dimensions plus représentatives des problématiques que nous développons dans le contexte des sciences de l'information et de la communication et dans celui des sciences humaines et sociales en général.

La *vélocité* (*velocity*) est l'une des dimensions à laquelle le travail pour PRÉDICTYS nous a confrontés. Cette dimension exprime la pression du temps soit que les données s'actualisent de manière fréquente, soit que l'utilisation des données s'effectue dans un temps réduit. La pression la plus forte intervient dans les situations de flux continu de données (*streaming*) appelant une analyse interprétative instantanée (*realtime*). Suivant les logiques d'exploitation, la production de résultats dans un temps déterminé peut imposer un ajustement de la performance impliquant les moyens et les techniques du *Big Data*. Le rapport au temps fait ainsi entrer tout un ensemble de problématiques, notamment liées au commerce électronique et à l'économie numérique, dans le champ du *Big Data*.

La *variété* est une autre dimension souvent évoquée conjointement aux deux précédentes (volume, vélocité, variété : les 3V du *Big Data*). Elle est associée à la diversité des ressources et des formats pouvant être impliqués dans les problématiques contemporaines. L'exploitation des données de

[228] Ce point de vue est distinct de celui plus global de la constitution du Web en tant que marché.

localisation dans un couplage avec la gestion de la relation client (CRM) est un exemple illustrant bien le double enjeu de la vélocité et de la variété[229].

De nombreux auteurs se contentent de ces définitions fondant la définition du *Big Data* sur une complexité computationnelle. D'autres auteurs, comme par exemple Pierre Delort, ne s'appuient que sur le couple volume et *densité* (Delort, 2015). La *densité* est un critère qualitatif d'évaluation des collections de données. Elle mesure la quantité de données non évaluées (absentes) dans la structure de données réalisant l'inscription. En considérant la collection comme une matrice d'enregistrements comportant plusieurs champs (valeurs), la densité, le creux, c'est-à-dire le ratio de cellules vides pour la matrice. Il nous semble possible d'associer à ce critère binaire (présence ou absence de valeur), la perspective continue qui apparaît avec la notion de véracité (*veracity*) ou confiance des données (probabilité). Ces différents éléments d'appréciation continue de la qualité nous paraissent pouvoir étendre le critère de densité sans en altérer la portée heuristique. Ces hypothèses, sont nécessaires dans le cas de données déclaratives faiblement contrôlées ou contrôlables extraites du Web ou provenant de conditions de capture dégradées. Cette nuance qualitative étend largement la perspective computationnelle et la rend plus pertinente dans le contexte des sciences humaines et sociales.

En suivant ces hypothèses, le *Big Data* correspond à des collections de faible densité pour un volume élevé. La raison évoquée est qu'une collection dont la densité est totale peut être abordée dans la continuité d'un paradigme classique de traitement, au bénéfice d'un effort méthodologique, technique ou technologique.

Par ailleurs, la variété des ressources informationnelles existantes basées sur les protocoles de l'Internet ouvre des perspectives vertigineuses dès lors que les données qui les constituent sont reliées (*i.e.* mises en relation) et accessibles. On peut espérer que dans l'avenir, les logiques du *linked data* se généralisent au sein du Web pour des données (quelle qu'en soit l'origine) ayant un caractère public voire pour celles qui font l'objet de publicisation au travers d'interfaces accessibles à tous. Cette perspective, associée à la constitution d'un *espace de publication* des données devrait aussi intéresser les acteurs de l'exploitation commerciale des données dans un usage privé. En effet, on observe déjà des mises en commun de données entre des acteurs qui interagissent sur la même cible (*prospects*). Par ailleurs, des acteurs intermédiaires assurant des médiations dans ce sens existent déjà (Ref.17). Ce type de collaborations circonstancielles (voire opportunes), fondées sur des alliances éphémères le temps d'une campagne, est susceptible de se développer. Les technologies du *linked data* constituent une étape supplémentaire dans l'interopérabilité des données. Le coût important de leur mise en œuvre est un frein non négligeable qui, cependant, peut être atténué par la mobilisation des abonnés et des internautes (cf. *digital labour*). Ce coût pourrait être compensé par les opportunités de valorisation que la mise en relations permet de concevoir.

Plus globalement et au-delà de la distinction des espaces marchands et non marchands, le développement de l'approche *linked data* du Web des donnés est tributaire de l'évolution et de

[229] Produire une offre commerciale à un client, en fonction de sa localisation dans un espace marchand en période de fin d'année ou de soldes est un cas d'étude.

l'appropriation des techniques analytiques par les acteurs. Le débat actuel entre *causalité* et *corrélation* témoigne de ces enjeux. La mise en œuvre d'un niveau sémantique porté par des relations soutient la voie inférencielle et causale. Elle va dans le sens d'une approche compréhensive des phénomènes observés mais constitue une voie longue. Or, l'incertitude et la pression du temps dans nos sociétés contemporaines dont témoignent de nombreux auteurs (Boyd, Crawford, 2011) conduisent à privilégier un temps court. Les approches fondées sur la recherche de corrélation sont de ce fait plus appropriées. Le développement actuel de l'ingénierie analytique et le positionnement d'acteurs puissants du Web sur ces sujets en sont la preuve. Les modèles fondés la recherche de régularités font florès, d'autant plus qu'ils sont de nature prédictive. Il va sans dire que tout modèle d'analyse ayant une potentialité prédictive - même si elle est accompagnée de réserves y compris d'ordre méthodologique - recevra un accueil très favorable auprès de décideurs confrontés à des choix stratégiques.

Rien n'est cependant figé. Les logiques d'usage ne sont pas installées, le marché ainsi que la réflexion sont encore balbutiants.

Quoi qu'il en soit, il parait nécessaire d'intégrer dans nos raisonnements le fait qu'on est amené à manipuler des données et des relations très partielles et imparfaites. On rejoint ici la caractéristique de *densité* évoquée précédemment. Ainsi caractérisé, le *Big Data* apparaît davantage comme le cadre d'une transformation profonde dans la manière de concevoir le rapport à la connaissance. Les questions soulevées portent sur la capacité heuristique dont on dispose à partir de collections de faible densité mais suffisamment volumineuses pour être productives suivant des règles inférentielles qu'il serait dommage de réduire aux seules règles déductives. Dans la perspective de constitution de collections, l'objectif poursuivi est d'étendre les possibilités heuristiques. Cependant, ce questionnement ne peut effectivement se passer d'une réflexion épistémologique sur l'actualisation des cadres de pensée vis-à-vis de la production de connaissances. C'est alors un tout autre débat qui est ouvert et sur lequel nous reviendrons au chapitre suivant.

3.2. Enjeux méthodologiques liés à la production de connaissances

Pour cette partie isolée de la production des données, la réflexion méthodologique est portée par les spécificités que nous avons associées au *Big Data* (densité, variété, vélocité). Ces critères sont au cœur de notre démarche, davantage que ne l'est l'augmentation de volume. Certes, les volumes importants conduisent aux limites certains algorithmes (et logiciels) ou techniques analytiques qui ne sont pas nécessairement adaptés ou éprouvés pour de (très) grands volumes de données. Mais ce point ne nous parait pas introduire de rupture méthodologique majeure, ce qui en revanche n'est pas le cas pour les critères précédents qui imposent d'autres approches.

C'est tout particulièrement le cas pour la densité qui rompt avec l'hypothèse habituelle de complétude appliqué aux données. La première des conséquences a été d'étendre les modèles de représentation des données aux modèles *NoSQL*[230] afin de traiter de manière homogène

[230] *Not only SQL*.

l'incomplétude et la diversité des structures récoltées. Ces modèles spécialisent le rapport aux données en fonction de leur nature (spatiale, graphe, documentaire, etc.) et des méthodes de traitements envisagées (*Map Reduce*[231], etc.). L'apport du modèle documentaire (MongoDB[232]) introduit une continuité entre document et données. Il donne aussi une à la notion de collection de données un aspect plus documentaire. Nous n'aurions probablement pas eu la même tendance à documenter les données en utilisant le modèle relationnel. Ainsi, ces modèles font évoluer la manière d'appréhender et de raisonner sur les données.

Plus globalement, les techniques de segmentation (*clustering*) et de projection (ACP[233], AFC, etc.) issues de l'analyse multidimensionnelle (cf. chapitre 3) sont adaptées à ces situations d'incomplétude des données et suscitent beaucoup d'intérêts. La réalisation de ce type d'analyse, souvent portée par des visualisations (plans factoriels, dendrogramme, etc.) encourage l'interprétation intuitive. Les agencements graphiques sont parfois sensibles à des paramétrages contextuels indépendants de la méthode. Il faut donc se méfier d'une lecture potentiellement biaisée par l'apparence graphique. Toutefois, la portée de ce type d'analyse est indéniable dans la mesure où elle est contrôlée et éprouvée avec application. Dans un contexte de forte variété, ces méthodes peuvent également faire apparaître des corrélations entre des données. À nouveau, il convient de se prémunir d'interprétations simplistes de ce qui se donne à voir. En revanche, il serait préjudiciable dans un contexte exploratoire de se priver de méthodes fortement heuristiques.

Les remarques précédentes formulées vis-à-vis de méthodes s'étendent aux logiciels qui les mettent en œuvre suivant des hypothèses et des choix techniques ou des paramétrages qui sont opaques pour l'analyste. L'effet boite noire lié aux interfaces des logiciels peut conduire là encore à des interprétations trop rapides des observations. De manière générale, la maîtrise de ces logiciels devrait s'accompagner de la maîtrise des méthodes et des techniques d'analyse des données. La collaboration avec des chercheurs du domaine des mathématiques appliquées et des statistiques a été pour nous une manière de pallier ces difficultés, tout au moins dans une période transitoire d'appropriation collective des enjeux du *Big Data* et de ses techniques analytiques. Il nous parait tout aussi important de gagner en compétences personnelles que de développer les interfaces disciplinaires. Au-delà des compétences et des tactiques individuelles, le débat qui s'ouvre alors est celui de la limite de l'interdisciplinarité. La maîtrise des outils et des méthodes numériques apporte un potentiel heuristique indéniable dont il serait dommage de se priver.

Le critère de vélocité quant à lui, nous parait essentiel car il rejoint l'injonction faite aux Sciences sociales de produire des analyses dans des délais extrêmement courts, au fil des événements médiatiques. Nous définissons ce nouveau paradigme analytique de *juste à temps* Cette contrainte temporelle, dans un contexte de flux massifs de données nous conduit à privilégier les modèles analytiques pouvant s'exprimer dans des algorithmes itératifs dont les caractéristiques de convergence permettent des interprétations au plus tôt.

---

[231] Schéma d'organisation de traitements en vue de leur parallélisation;

[232] https://www.mongodb.com

[233] Analyse en Composant Principal, Analyse Factorielle des Composants.

Si la recherche académique n'a pas à se plier aux injonctions du temps réel, cette prise en compte nous parait justifiée pour deux raisons. Tout d'abord parce qu'il s'agit d'une condition nécessaire pour certains type de traitement impliqués dans le *monitoring* ou les processus d'analyse diachronique. Cette contrainte se traduit par une linéarisation temporelle des représentations au sein de la collection. Enfin, dans une posture critique, il nous parait important d'analyser les conditions et la faisabilité d'une analyse en temps réel.

Pour conclure, même si la performance des procédés d'observation fait l'objet d'un travail approfondi, les conditions de leur réalisation ne garantissent ni l'exhaustivité, ni la représentativité des collections, voire l'objectivité. La mise en œuvre de techniques massives ne vise pas l'objectif de satisfaire à l'un ou à l'autre de ces critères. Comme le souligne Dominique Boullier, il est nécessaire de dépasser ces critères et ne pas s'arrêter aux limites qu'ils font apparaître (Boullier, 2015, p3). Ce dépassement ne signifie pas ignorer les limites mais donner une autre portée au travail d'analyse. La réintroduction de la variable temporelle est de notre point de vue une clef de cette réorientation. En effet, l'observation porte sur des plages continues, dans lesquelles surgissent et se développent des phénomènes remarquables que d'autres techniques ne permettent pas de saisir dans leur dynamique. Ces phénomènes peuvent être sporadiques ou éphémères et correspondre à une instabilité remarquable sous certaines conditions. Fixer ces éléments dans des collections est une opportunité d'analyse pouvant enrichir d'autres résultats. C'est au travers de la nouveauté de ces observations qu'il convient d'étendre le champ d'investigation et le registre d'analyse. Cette approche ne fait pas table de rase des autres méthodes : elle les complète.

CHAPITRE 6
Enjeux Scientifiques et Disciplinaires

« L'avantage des nouvelles méthodes
[du numérique] est qu'elles permettent de tracer
l'assemblage des phénomènes collectifs
au lieu de les obtenir par agrégation statistique. »
(Venturini, Latour, 2015)

La dimension méthodologique évoquée dans les deux précédents chapitres contribue aux transformations associées à l'évolution des pratiques scientifiques en Sciences humaines et sociales et en Sciences de l'information et de la communication. Ces éléments méthodologiques issus de notre réflexion sont associés à une démarche empirique et expérimentale conduite dans le cadre d'études se rapportant aux médiations instrumentées. Les technologies (dont celles d'Internet et du Web), constitutives des dispositifs info-communicationnels numériques[234], ont un rôle qui dépasse les médiatisations et les services qu'elles permettent d'accomplir. Elles servent d'autres finalités, relatives aux dispositifs, telles que la sécurisation et la production de valeur, en assurant conjointement des fonctions de traçabilité ou de traçage de plus en plus fines. Ces fonctions, déjà anciennes, ont acquis une place centrale dans la mise en œuvre des dispositifs en raison de leurs implications dans la chaîne de valeur et l'économie des données.

La centralité de la trace provient de deux nécessités que nous associons au processus de *datafication*[235]. Celle d'une part, de production d'information-valeur ou *informationnalisation* que l'on trouve évoquée dès le début des années 1990 dans une approche renouvelée de l'économie de l'information, suivant une tradition qui est plutôt celle des Sciences de l'information[236]. Celle d'autre part, tout aussi forte, de *computation* que nous (re)définissons dans un double emprunt : à sa définition française de *manière de calculer le temps* et à sa définition anglaise très étroitement liée au calcul et à l'informatique. Nous avons vu que ces nécessités ont été incorporées dans les technologies constitutives des dispositifs info-communicationnels lorsque ceux-ci requièrent une traçabilité individualisée répondant au besoin d'*accountability*. L'extension du principe d'*accountability* a déplacé le focus de la traçabilité vers le traçage individualisé des interactions. Désormais la majorité des fournisseurs de services organisent des bases de données à partir de ces traces dans la perspective de valoriser la connaissance qu'ils détiennent sur leurs abonnés.

---

[234] Désignés par la suite comme dispositifs.

[235] Notion précédemment évoqué au chapitre 3

[236] Voir par exemple la critique de Jean-Michel Salaün du livre de Thierry Ribault : *l'économie de l'information – approche patrimoniale* parue aux éditions A jour en 1993. http://bbf.enssib.fr/consulter/bbf-1994-03-0089-023

Formellement, la computation caractérise une classe plus générale de traitements qui contient la classe réunissant la traçabilité et le traçage. Cette généralisation conceptuelle permet d'intégrer l'ensemble des traitements effectifs et potentiels qui peuvent être enchaînés à la suite de la capture numérique des interactions dans les dispositifs.

Pour faire suite à la présentation de nos travaux, il s'agit désormais de mettre en perspective les résultats auxquels nous aboutissons sur les plans théoriques, épistémologiques et méthodologiques. Au-delà de l'évaluation de notre contribution au domaine d'étude des médiations info-communicationnelles, l'objectif est d'envisager comment nos travaux, très clairement tournés vers le numérique, nous permettent d'aborder le devenir des pratiques sociales et techniques, et de quelle manière ils contribuent au renouvellement des questionnements qui fondent les Sciences de l'information et de la communication.

Compte tenu de la position particulière de nos travaux dans le champ des SIC, il nous parait utile de revenir sur deux notions centrales mises en avant dans ce mémoire. En premier lieu, les TIC, jusqu'ici considérées comme des faits techniques isolés se fondent maintenant dans des dispositifs complexes dont les configurations varient dans le temps. Il faut donc adopter une approche différente intégrant les spécificités des dispositifs. En second lieu, la production de données (ou *datafication*) en tant que processus technique doit être intégrée au cadre théorique des SIC. Ce processus doit être pris en compte dans la modélisation de l'*informationnalisation* comme dans l'approche empirique des médiations, ce qui n'est pas le cas jusqu'à présent.

1. La notion de TIC dans les Sciences de l'information et de la communication

Dès l'origine de la discipline, son assise s'est établie dans une proximité étroite avec la technique et les technologies de l'information. L'importance de la cybernétique et de l'informatique dans l'épistémologie des Sciences de l'information et de la communication est régulièrement rappelée. Ainsi, le numérique apparaît comme un objet ou un champ de transformations socio-techniques dont l'étude est légitime du point de vue disciplinaire. Il s'agit d'une légitimité construite par rapport à une interdisciplinarité dans un rôle qui devrait être celui «*…de "traducteur" entre les humanités et les technologues du numérique,…*» comme l'évoquent Imad Saleh et Hakim Hachour (Saleh, Hachour, 2012, p7).

Cette position d'entre-deux est effectivement occupée depuis plus de 30 ans par les SIC comme le révèle l'étude des publications disciplinaires menée par ces deux auteurs (*Ibid.*). Mais la diversité et la distribution étendue des problématiques abordées ne suffisent pas. La position occupée sur le territoire conceptuel n'est qu'une première étape de la traduction. Le rôle de *traducteur* suppose une connaissance approfondie des technologies numériques qui va au-delà d'une rhétorique maîtrisée. Cette condition ne nous parait pas uniformément remplie. Les Technologies de l'Information Communication (TIC) notamment semblent fonctionner comme un substantif intemporel alors qu'il est pourtant dit qu'elles deviennent numériques et qu'elles se renouvellent en permanence. Ainsi, alors que les technologies et la technique sont mobilisées dans l'épistémologie de la discipline,

leur connaissance ne progresse que très peu, au risque d'entretenir à leur sujet, des discours périphériques ou de surface.

1.1. Une définition multiple

La définition du terme *technologie* à laquelle il est fait référence dans cette expression correspond de manière consensuelle à la traduction anglaise actuelle du terme *technology* : « *refers to methods, systems, and devices which are the result of scientific knowledge being used for practical purposes* »[237]. Il s'agit d'une définition pratique, éloignée de celle d'étude des techniques.

Considérée au singulier, *une TIC* désigne généralement une réalisation technologique particulière. Pour Alex Mucchielli, une TIC ne peut pas désigner une forme abstraite (type). Elle doit être identifiée à une réalisation (instance), c'est-à-dire «*…insérée dans un dispositif social, donnant lieu à des usages »*[238] (Mucchielli, 2006, p21). Cette approche d'une instance technologique, toujours liée à un contexte spécifique, n'interdit pas de considérer dans le même temps l'existence d'une forme plurielle. Dans ce cas, *les TIC* désignent l'accumulation de toutes les déclinaisons possibles de réalisations technologiques constituant une classe. Le cadre de cette proposition de définition nous parait trop contraint pour donner une étendue suffisante du champ technologique et de ses dynamiques. Il nous semble préférable d'adopter une définition plus ouverte et théorique. Cette définition peut se construire directement sur la forme plurielle des technologies de l'information et de la communication. Il s'agit alors d'une définition en intension.

Dans ce cadre, l'acronyme pluriel TIC agrège l'ensemble des connaissances qui expliquent les processus et les logiques industrielles, socio-économiques et culturelles associées au cycle de vie des technologies et des techniques qui sont à l'origine des processus info-communicationnels médiés au sein d'une communauté organisée. Il est probable que cette tentative de définition soit elle-même imparfaite. Il s'agit en effet d'une expression valise bien commode pour englober et désigner tout ce qui assure la disponibilité, la continuité et l'accomplissement des processus info-communicationnels médiés. Bien que la dimension temporelle soit implicite, il convient de préciser que cette définition est relative à une période donnée. Mis à part les discours sur les *Nouvelles* technologies (NTIC) qui sur-définissent l'actualité technologique et sa capacité au renouvellement et à l'innovation, le positionnement temporel des TIC est le plus souvent imprécis. Cette imprécision confère alors à l'expression une élasticité propice à la généralisation des discours. Notons cependant, que le cycle de vie des technologies en général n'est pas nécessairement le même que celui des objets qu'elles produisent, ni celui des usages de ces objets.

Dans le même temps, la référence généralisée aux technologies est nécessaire à la construction d'un discours scientifique. Cependant, la réduction des artefacts (objets, machines, services) à des

---

[237] Extrait du dictionnaire COBUILD édité par Collins. Version en ligne consultée (01/09/2016) : http://www.collinsdictionary.com/dictionary/english/technology

[238] Dans le texte dont cette citation est extraite, Mucchielli réduit les TIC aux dérivés informatiques, ce qui nous semble excessif.

technologies identifiées n'est que très rarement réalisée : les technologies sont amalgamées ou résumées à celles qui apparaissent dominantes.

Par ailleurs, l'estampillage Info-communicationnel (IC) apposé aux technologies isole un sous-ensemble spécialisé. Cette spécialisation est double. Elle correspond à une finalité si l'artefact sert à informer ou à communiquer. Elle correspond à une qualité si l'artefact communique lui-même pour accomplir sa fonction. Cette double lecture du fait technologique contribue à la coexistence de deux types de déterminations, sociale ou technique, dans la construction des artefacts. Cette double polarisation est positive dans une approche réconciliée de type socio-technique. Cependant, elle est toujours l'occasion, au sein de la discipline, d'une mise à distance de la technologie.

1.2. Un impensé disciplinaire ?

Ces imprécisions et ambiguïtés qui entourent la définition des TIC sont régulièrement relevées dans les publications (notamment en SIC) mais ne conduisent pas à des propositions alternatives. Les TIC constituent une catégorie admise, et en apparence stable, alors que les technologies et les modalités de leur usage évoluent. Les TIC sont-elles un *impensé*[239] de notre discipline, alors qu'il est régulièrement réaffirmé qu'elles en constituent l'une des clefs fondatrices ?

De notre point de vue, le premier élément explicatif de l'utilisation de la notion de TIC réside dans la caractéristique des technologies à s'organiser en système. Nous rejoignons en cela l'explication des objets techniques donnée par Gilbert Simondon dans le chapitre qu'il consacre au processus de *concrétisation* dans l'ouvrage *Du mode d'existence des objets techniques* (Simondon, 1858). Ce processus est présenté comme celui par lequel l'objet technique acquiert sa consistance tout en convergeant vers une forme stabilisée, palier provisoire dans son devenir d'objet[240] (*Ibid.*, p20). Cette proposition, que nous acceptons, nous permet de justifier l'utilisation de la notion de TIC. Elle rejoint la proposition que nous faisons d'une définition des TIC en intension.

On peut ainsi considérer que le recours au terme abrégé TIC est une manière de se dégager de contingences qui s'exprimeraient plus fortement sur les artefacts que sur les technologies.

Par ailleurs, l'emploi du terme TIC souligne les interdépendances socio-techniques nécessaires à la concrétisation d'artefacts. On retrouve dans ces considérations l'accent privilégié mis historiquement en SIC sur les conditions de production associées aux industries culturelles ou médiatiques.

Dans ce contexte de définition globale, il arrive que des auteurs[241] utilisent la forme singulière d'une TIC pour désigner une branche particulière du réseau complexe des TIC et traduire ainsi sa domination ou sa relative autonomie.

Cette lecture orientée par le processus de composition (production) doit être complétée d'une seconde lecture dirigée à partir de l'instrumentation des médiations vers les artefacts. Cette seconde

---

[239] Voir Une théorie sociétale des TIC : Penser les TIC entre approche critique et modélisation conceptuelle Pascal Robert.

[240] L'objet technique n'est jamais figé.

[241] B. Miège notamment.

approche n'est pas contenue à l'origine dans l'expression TIC. Elle est toutefois plus conforme aux logiques d'usages qui ouvrent la possibilité d'étudier des artefacts détournés d'une finalité première qui ne serait pas de nature info-communicationnelle.

Cette extension du domaine des TIC introduit une formulation intentionnelle plus lâche. Ainsi, nous pouvons considérer comme TIC des systèmes technologiques sans que chacune des technologies qui les constituent aient nécessairement un caractère info-communicationnel. Par ailleurs, le caractère numérique[242] ne s'appose à un artefact ou à un système que si l'une au moins des technologies utilisées pour le produire dispose de cette qualité. En conséquence des principes de la cybernétique, tout artefact ou système numérique appartient aux TIC. Parmi les technologies numériques, celles de l'Internet (et du Web) occupent une place centrale dans les préoccupations actuelles des chercheurs en SIC au point qu'il y a quasiment assimilation avec les TIC.

Pour revenir sur la question de l'*impensé*, le problème qui se pose est propre à un objet frontière. Les Technologies de l'information et de la communication ont identifié des artefacts circonscrits dans leur matérialité et leurs finalités comme dans leurs logiques de production. La connaissance de surface que l'on pouvait avoir des appareils suffisait à établir leur unité technologique et les réseaux d'acteurs qui leur étaient associés. L'essor des technologies numériques a transformé ce mode d'existence en l'étendant à des formes plus diffuses et réticulées. L'unité de façade ne permet plus de saisir les artefacts de manière satisfaisante. Des connaissances techniques sont nécessaires, supposant parfois le déplacement de frontières disciplinaires, notamment informatiques. Il ne s'agit pas d'un impensé mais de l'épuisement d'une notion qui se vide de sa substance.

Évoquer les TIC suivant un niveau de granularité de plus en plus spécialisé, permet-il encore de restituer la réalité des pratiques et des usages sociaux ? Dans la publication de 2007, du tome 3 de la série *La société conquise par la communication*, B. Miège écrit : « *Les dispositifs sont/seront des configurations sociotechniques appelées à assurer le développement des Tic sur la durée et à donner des bases renforcées à la médiation technique de la communication […] l'on ne saurait désormais envisager la question des déterminations techniques outil par outil : le dispositif est une configuration technique à appréhender en tant que telle*… » (Miège, 2007, p48).

Dix ans plus tard, l'évolution que cet auteur estimait prématurée doit être accomplie. Le concept de dispositif info-communicationnel nous parait indispensable comme cadre d'analyse. Ce repositionnement conceptuel intégrant la complexité d'un dispositif modifie le rapport aux déterminations socio-techniques qui prévalaient jusqu'alors dans l'analyse des pratiques.

2. La *datafication* à l'épreuve des SIC

La réflexion que nous avons menée jusqu'à présent a été guidée par la préoccupation expérimentale de l'observation écologique. Au moyen d'un dispositif adapté, il s'agit de produire des données qui traduisent le fonctionnement situé du dispositif observé[243]. L'activité d'observation vise alors la

[242] Les technologies numériques contiennent les technologies informatiques

[243] Qui inclus dans notre définition les utilisateurs.

production de collections structurées répondant à des objectifs larges, d'analyse et de communication scientifique.

Pour produire de la connaissance, nous devons mettre en regard les deux versants de la production de données : celui des dispositifs observés et celui des dispositifs observateurs. Bien que l'objectif soit de rapprocher ces ensembles de données, il existe un écart théorique entre les deux systèmes de représentation. Pour que cette confrontation puisse servir une réflexion scientifique, il est nécessaire de recourir à un cadre théorique pour situer et évaluer la production de connaissances. C'est la raison pour laquelle nous abordons la notion d'*informationnalisation* tel qu'il apparaît dans les SIC. Ce concept nous permet de proposer une modélisation des processus de production de données relativement symétrique entre les deux familles de dispositifs.

2.1. L'*informationnalisation* comme processus

L'*informationnalisation* telle qu'elle est envisagée au tournant des années 2000 par Bernard Miège et Gaëtan Tremblay (Miège, Tremblay, 1999) caractérise la capacité humaine à engendrer ce que nous avons identifié comme des dispositifs info-communicationnels. Il s'agit selon B. Miège « *d'un procès ou d'une logique sociale de la communication qui se caractérise par la circulation croissante et accélérée de flux d'information éditée ou non, autant dans la sphère privative, dans celle du travail que dans l'espace public* » (Miège, 2007, p66). La nature de dispositif est attestée par les phénomènes de domination qu'évoque l'auteur et qu'il associe à « *la possession, la disposition et le maniement* » des outils (*Ibid.*, p67). Ce processus est défini par B. Miège comme un phénomène, dont les flux d'information sont les manifestations. La notion de flux fait référence à la délimitation d'un contenu ainsi qu'à sa mise en circulation. C'est en particulier dans le fait qu'elle circule en affectant des individus qu'une donnée deviendra information. Cette définition générale de l'*informationnalisation* englobe celle que nous évoquions en introduction de ce chapitre. Elle est cependant réduite aux faits de communication survenant dans l'espace social.

Les flux d'informations éditorialisées peuvent être assimilés aux flux médiatisés qui irriguent la partie visible du processus d'*informationnalisation* affectant notamment l'espace public. La partie complémentaire, non éditée, correspond aux productions informationnelles dérivées, associées aux différentes régulations et contrôles nécessaires à l'accomplissement du processus d'*informationnalisation*. L'approche que nous poursuivons nous incite à préciser davantage cette partie en mettant l'accent sur les dynamiques socio-techniques du dispositif à l'œuvre. Notons que la distinction productive/non productive[244] de l'information que B. Miège rappelle (*Ibid.*, p69) ne dépend ni de la nature du dispositif ni du processus d'*informationnalisation*.

2.1.1. La *datafication* et la computation

Il nous parait opportun d'affiner le processus d'*informationnalisation* en soulignant l'importance d'un flux dérivé que nous associons aux processus de *datafication* et de *computation*. Notre hypothèse est

[244] La distinction n'est pas aisée. Elle s'appuie sur une stabilité institutionnelle durable ou éphémère associée à une économie de l'information.

que la finalité de ces flux d'information n'est pas seulement d'assurer le procès global de communication mais aussi de capturer ou créer de la valeur dans la production et l'enrichissement de données, ce qui apparaissait dans la première définition de l'*informationnalisation*.

La notion de *datafication* a été popularisée à partir de 2013 par Kenneth Cukier et Victor Mayer-Shoenberger (Cukier, Mayer-Schoenberger, 2013, 2014). Le processus qu'il désigne est distinct de celui de numérisation (*digitization*) qui porte sur la transformation de contenus analogiques préexistants. Pour ces auteurs, la *datafication* consiste à produire des données numériques à partir du *monde réel* pour en quantifier certains aspects. Ils resserrent cette définition aux aspects de la vie quotidienne[245] délimitant ainsi le champ d'études et d'applications aux Sciences humaines et sociales. La définition de la *datafication* est associée au processus de *digitalization*, c'est-à-dire d'utilisation du numérique comme technologie élémentaire. La notion de *datafication* est proposée pour expliquer un cadre particulier de l'augmentation, sur l'Internet global, du volume de données numériques consacrées aux services commerciaux ou marchands. Ce terme est mobilisé dans des travaux concernant une analyse macroscopique de l'économie de l'information et des données. Il s'applique aux stratégies développées et aux inflexions apportées dans ces secteurs par les (très) grands acteurs dominants du Web[246].

Nous prolongeons cette notion au niveau micro en lui donnant un sens opérationnel. Par ailleurs, la proposition de définition de K. Cukier et V. Mayer-Schoenberger nous semble devoir être revisitée : la production de données quantifiées (*quantified data*) est une définition trop restrictive[247]. Cette définition met un accent trop important sur la mesure et l'objectivation. Elle correspond cependant à une classe de travaux comme, par exemple, les travaux portant sur le phénomène de la mesure de soi-même (*quantified self*).

Cette définition de la *datafication*, amenant l'usage de métriques, est orientée vers une définition du paradigme de performance à finalité évaluative ou décisionnaire. Cette visée s'adresse prioritairement au contexte productif[248] qui n'est plus l'unique finalité de la production des données. La *datafication* doit exprimer la production de données dans toute sa généralité, y compris celle des données qualitatives (catégorielles), objectives ou subjectives. Cette ouverture de la définition permet d'utiliser ce terme dans le contexte de l'analyse exploratoire propre à la démarche scientifique.

Nous considérons la *datafication* comme le processus de production de données numériques associé au processus d'*informationnalisation* du dispositif. La *datafication* s'appuie sur les flux informationnels

---

[245] « *Datafication is a far broader activity: taking all aspects of life and turning them into data* » (Cukier, Mayer-Schoenberger, 2013, p6). Nous traduisons *life* par "vie quotidienne" qui nous parait adapté pour soutenir la référence à l'existence individuelle portée par des rythmes, des contraintes, etc.

[246] Souvent résumé depuis 2013 dans l'acronyme GAFA(M) : Google Apple Facebook Amazon et Microsoft (ce dernier étant parfois oublié).

[247] D'où les critiques, notamment de *dataism* (Van Dijck, 2014) qui peuvent être adressées à leurs travaux.

[248] Au sens d'une industrialisation.

du dispositif. Ces flux peuvent soit être produits par des ressources du dispositif, soit servir une finalité opérationnelle ou stratégique pour des *ressources*[249] ou pour le dispositif lui-même.

On distingue deux grands types d'organisation de données issues de ces flux : les organisations traditionnelles qui correspondent à l'évolution continue des systèmes d'informations classiques ; les organisations répondant au contexte des *Big Data* qui nécessitent une infrastructure, des techniques et des compétences différentes. Nous désignons par computation les adaptations du processus de *datafication* qui sont nécessaires dans le contexte des *Big Data*. La computation, qui requiert des compétences expertes dans le traitement et l'exploitation de données, peut être réalisée par une ressource dédiée à l'intérieur du dispositif.

La formalisation inhérente au processus de *datafication* permet de distinguer les aspects représentationnels qui, d'une part, sont fonctionnellement nécessaires à la ressource et la caractérisent d'un point de vue externe, de ceux qui, d'autre part, vont au-delà de l'*informationnalisation* et servent la computation.

En associant la *datafication* à l'*informationnalisation*, nous avons assujetti la donnée à l'information. Suivant cette perspective, la *datafication* est poussée (*push*) par les flux d'informations. Or, rien n'interdit que les logiques de *datafication* ou de computation soient dominantes, voire à l'origine de l'interaction ; autrement dit, le dispositif peut être ponctuellement asservi par des logiques représentationnelles devenues structurantes de son activité. Nous considérons ce second régime de fonctionnement dirigé comme tiré (*pull*) par la nécessité des données. Ce régime de fonctionnement plus ou moins transparent, est imposé *a priori* aux usagers. Au bénéfice d'ajustements et de concessions visant la normalisation, il tend à s'intégrer aux logiques d'usages. L'authentification personnelle est un exemple parmi d'autres de ce régime de fonctionnement, le bouton *like* en est une autre illustration.

Il serait illusoire de considérer que ces régimes de fonctionnement sont neutres vis-à-vis de l'usage du dispositif. Dans ce contexte de contrôles manifestes, il convient de mettre en regard un régime de participation individuelle faisant apparaître différentes modalités. Ces modalités traduisent les formes d'engagement des usagers-acteurs, plus ou moins consenties, plus ou moins spontanées, plus ou moins transfigurées, etc. Les différents accords possibles de ces modalités dans la participation individuelle jouent un rôle fondamental. Les caractéristiques quantitatives et qualitatives des données issues de la *datafication* en découlent et conditionnent la computation.

2.1.2. Éditorialisation des données

Nous introduisons un processus d'*éditorialisation des données* afin d'exprimer les contrôles exercés sur les données lors de leur mise à disposition dans les interfaces publiques (API) ou dans les publications qui les utilisent (Web). Le terme *éditorialisation* est surtout utilisé depuis les années 1990 pour désigner l'adaptation des processus éditoriaux aux conditions spécifiques d'un support.

[249] Nous sommes amené à distinguer les deux niveaux de ressource (entité) et de dispositif (tout) dans une définition récurrente, une ressource pouvant elle-même être à son niveau un dispositif. Soulignons que dans le cas d'observation, nous n'accédons potentiellement qu'aux ressources du dispositif observé (et non au dispositif lui-même, sauf s'il est réduit à une ressource).

L'éditorialisation traduit les ajustements techniques nécessaires pour investir un support d'objectifs éditoriaux et de publication. Cette notion est désormais envisagée pour décrire la cohérence des traces produites collectivement à l'occasion des interactions sur les réseaux sociaux qu'engendrent des événements (Merzeau, 2013). Dans ce cas, l'adaptation résulte de la dynamique socio-technique de l'usage.

Dans le cas présent, la sélection et les différentes opérations mises en œuvre, lors de publication ou lors d'instanciation de données publicisées ou publiées, sont assimilées à la conduite d'un projet de valorisation éditoriale. Ce projet est par nature contraint par les enjeux internes et stratégiques de la ressource publiante ainsi que par sa participation dans le dispositif selon la configuration courante. L'éditorialisation assure une fonction de transformation de nature pour la donnée. Cette transformation se double d'une fonction de coupure entre l'espace productif interne et l'extérieur.

Ainsi, la présence d'un tel processus se traduit par un écart formel existant entre le cœur représentationnel (des données) de la ressource et ce qu'elle rend accessible à ses bornes. Dans notre propos, la trace numérique d'usage est le résultat de l'éditorialisation. Elle est un habillage des données. En ce sens, le statut de document nous semble plus approprié que celui de donnée pour évoquer les traces numériques d'usage.

Il faut donc avoir à l'esprit qu'il peut exister un biais dans l'interprétation que l'on peut faire des données rendues accessibles au travers des traces.

Nous justifions la référence éditoriale en partant du constat du continuum qui s'est installé entre document et donnée. Cela tient autant : à l'émergence de formats structurés, à l'atomisation structurelle qui en découle, qu'à la *redocumentarisation* à l'œuvre dans l'espace numérique (Salaün, 2007). L'évocation d'un continuum provient principalement de considérations représentationnelles qui nous guident dans ce mémoire. De ce point de vue, données et documents s'inscrivent dans des traitements équivalents sans pour autant être assimilés l'un à l'autre. Il est cependant difficile de ne pas céder à la facilité du passage de l'un à l'autre dans une compréhension synecdotique.

Selon un axe de réflexion plus général, la référence éditoriale est aussi celle du modèle théorique de production issu des travaux sur les industries culturelles (Miège, 1996, pp177-186). Le continuum précédent sensibilise progressivement les industries éditoriales à la question de la redocumentarisation et donc de la donnée. Nous pouvons constater que cette question rapproche la nature des activités des secteurs du contenu avec celles du secteur des données par une hybridation des compétences en cours. Dans le même temps, les leaders des industries de la donnée (Google, Amazon, etc.) sont dans une situation d'élargissement de leur assise en assimilant, dans les regroupements qu'ils organisent, des activités éditoriales. Cependant, on ne peut toutefois pas considérer que les deux secteurs d'activités fusionnent.

Le point souligné par les travaux de Louise Merzeau, comme d'ailleurs nous l'avons constaté au sein de la société PRÉDICTYS, c'est l'intérêt pour des industriels de la donnée personnelle d'être partie prenante d'une activité de circulation de contenus en soutien de la captation d'informations et donc de production de données.

2.1.3. Modélisation des processus

Les définitions précédentes nous amènent à proposer une modélisation dont la formalisation répond à des logiques informationnelles. Elle est également d'inspiration cybernéticienne dans sa représentation. Dans cette modélisation nous faisons l'hypothèse d'une équivalence fonctionnelle entre les différentes ressources constituant le dispositif. Les singularités de chacune sont supposées s'exprimer dans des variables de configuration. Les régulations internes au dispositif sont réalisées dans le processus d'*informationnalisation* que nous prenons comme référence.

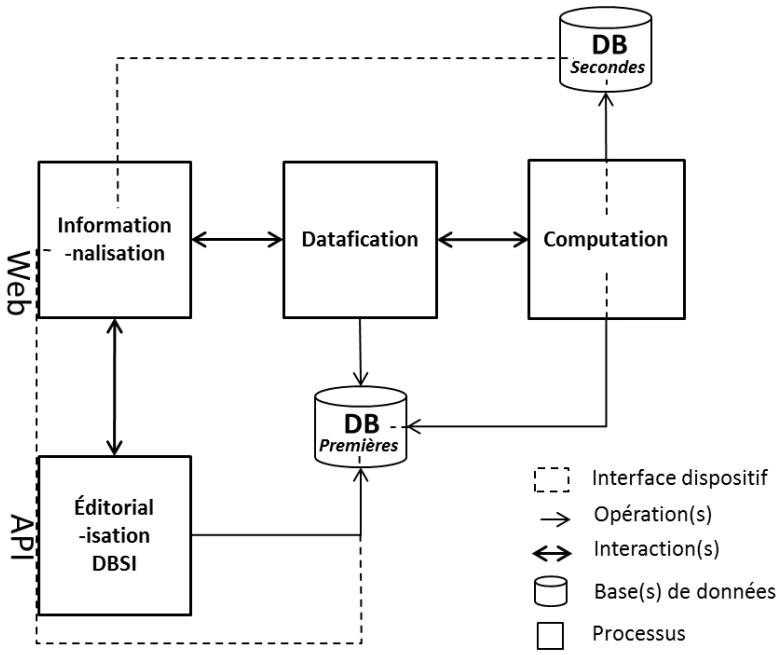

Fig39. Schématisation du processus d'informationnalisation propre à une ressource

Dans cette représentation, les limites de la ressource (tracées en pointillés) constituent une interface avec les ressources internes ou externes du dispositif.

Cette interface se décline de manière opérationnelle suivant différentes finalités. Par exemple, la mise en œuvre de services Web, la gestion d'une API, etc. Cette possibilité d'adaptation finalisée ou non, coordonnée ou non dans le cadre d'un dispositif, maitrisée ou non à l'échelon de la ressource, permet d'imaginer des propositions informationnelles très différentes suivant les bornes considérées (Web, API, etc.). Par hypothèse, le processus d'éditorialisation est considéré comme une propriété de la ressource et de ses enjeux. Suivant cette hypothèse, les productions aux bornes de l'interface deviennent le résultat d'ajustements entre les processus d'*informationnalisation* et d'éditorialisation. De manière analogue, nous considérons que les données premières correspondent à des ajustements entre le processus d'*informationnalisation* et le processus de *datafication*.

La place centrale occupée par les données dans cette formalisation traduit l'hypothèse d'une cohérence et d'une coordination entre les différents processus. Le processus de computation est envisagé dans un rôle complémentaire, comme le moyen d'introduire une perturbation, un

déséquilibre représentationnel porteur d'autres dynamiques, agissant par exemple sur la configuration du dispositif.

Soulignons que ce schéma peut aussi se lire dans le contexte de la définition des dispositifs d'observation, chacun des processus pouvant se transposer d'un dispositif à l'autre. La *datafication* expérimentale[250] correspond alors à la fabrication des données de collections assimilables dans le schéma aux données premières. L'éditorialisation des données et l'*informationnalisation* sont alors liées à la collaboration scientifique établie sur les données. La production de connaissance ne s'envisage qu'à un niveau plus abstrait, non décrit dans ce schéma simplificateur.

2.2. La *datafication* expérimentale au regard de la production de connaissances

Le cadre de notre réflexion est celui de la production de connaissances empiriques à partir de collections de données expérimentales. L'approche empirique n'est pas en soit le questionnement principal de nos travaux. Cette démarche de recherche, que nous avons adoptée dans nos travaux exploratoires, figure parmi les bonnes pratiques mises en avant et encouragée dans les manuels en SIC. Elle s'est développée dans notre discipline à la suite de l'ensemble des Sciences humaines et sociales à partir de la mise au point de méthodes quantitatives issues de l'analyse statistique au tournant des années 1960. Les SIC ont également accompagné le développement de méthodes d'analyses qualitatives selon une démarche empirico-inductive. Ce foisonnement méthodologique correspond aux fertilisations croisées de disciplines que les SIC ont naturellement accueillies en leur sein. Par ailleurs, la proximité des SIC avec des disciplines comme la psychologie, les sciences de gestion, (…) contribue à une familiarité avec les méthodes expérimentales.

La *datafication* sous-jacente à l'approche expérimentale nous incite à raisonner la production de connaissances au regard des représentations et de ce qu'elles autorisent ou interdisent.

2.2.1. Transformation et dégradation informationnelle

Le schéma suivant formalise l'enchainement des transformations opérant à partir du processus originel d'*informationnalisation* du dispositif observé et jusqu'au processus terminal de *datafication* expérimentale du dispositif d'observation visant la production d'une collection. L'objectif est désormais d'avoir une vue globale de la production de données dans le contexte de l'observation.

Pour le dispositif observé, les éléments de cette chaine de transformation informationnelle correspondent aux interfaces spécifiques de la ressource et du dispositif. Dans le cas présent, nous retenons deux interfaces API et Web services.

Pour les deux interfaces, nous faisons apparaitre deux composantes de transformation. Ce dédoublement vise à rappeler que les données disponibles aux bornes sont le résultat d'un processus d'éditorialisation d'une part, et du fonctionnement technique du service d'autre part. Cette double nature ne doit pas être oubliée dans l'interprétation des flux de données produits. Ainsi, le ratio d'un hashtag donné dans un flux instantané délivré par Twitter ne peut pas

[250] Le qualificatif "expérimental" permet de distinguer les deux contextes selon qu'ils portent sur les dispositifs observés ou les dispositifs d'observation.

s'interpréter comme la mesure de l'intérêt partagé pour cette thématique, sans évaluer ou formuler des hypothèses sur les effets de cadrages éditoriaux auxquels procède Twitter ainsi que sur les effets de filtrages que le fonctionnement technique induit.

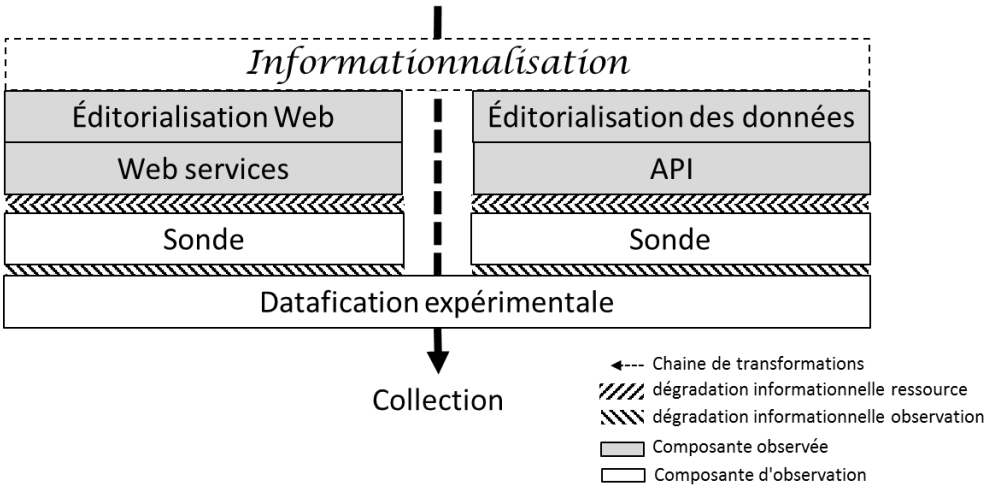

Fig40. Transformation et dégradation informationnelle

Il s'agit moins d'une mise en garde que d'une préconisation méthodologique qui conduit à étendre l'investigation et à croiser les résultats.

En plus des transformations identifiées, cette chaîne fait apparaitre des dégradations informationnelles qui relèvent des situations de fonctionnement respectif des deux dispositifs. Ces dégradations se traduisent par une dénaturation du flux informationnel en volume ou en qualité. Formellement, elles n'apparaissent qu'à l'interface des deux dispositifs et ne se manifestent que dans la production de données.

Le caractère autonome et médiateur de la sonde fait apparaitre une double dégradation potentielle sur les canaux la reliant à la ressource et au dispositif d'observation.

Dans le cadre expérimental, cette dégradation fait l'objet d'une objectivation ; il s'agit de ce fait d'une dégradation perçue qui est rapportée dans la collection. Certains aspects de cette dégradation tiennent aux configurations internes conjoncturelles ou permanentes. En l'absence d'indicateurs explicites, les configurations sont supposées installées et stables dans la durée de l'observation.

Dans le cas de la ressource d'observation, la configuration fait l'objet d'une optimisation afin de réduire les effets de dégradation perçue. D'autres aspects de la dégradation sont des aléas contingents. Ils proviennent de causes externes à la ressource ou au dispositif d'observation et constituent une incertitude qu'il conviendrait de pouvoir évaluer.

2.2.2. La collection, charnière indispensable

Ce sont des considérations méthodologiques qui nous ont amené à faire exister l'objet représentationnel qu'est la collection de manière distincte des corpus. Du point de vue représentationnel, la collection est ainsi plus éloignée des contraintes analytiques et plus proche des spécificités des ressources de l'Internet et du Web. La collection favorise l'exploration de l'espace

de données en créant un point de reprise (*backtracking*[251]) dans la constitution de corpus. Cette facilité associée à une approche incrémentale de la collecte peut aussi s'avérer utile du point de vue de la production des connaissances.

En outre, l'indépendance de la collection vis-à-vis des outils, des techniques et des méthodes d'analyse nous parait avantageuse. L'enregistrement, c'est-à-dire l'unité élémentaire de la collection, n'est pas une structure de données contrainte comme le sont les enregistrements de corpus finalisés. Les choix représentationnels portant sur l'enregistrement sont de ce fait plus ouverts et les contraintes associées à la publicisation sont sans effet sur les choix informationnels. En étant moins focalisé sur des spécificités analytiques de contenus, le caractère sériel ou temporel de la collection est ainsi plus affirmé.

De ce fait, la recherche de régularités structurelles (motifs) et de phénomènes temporels permet d'établir des cadres pour l'analyse globale du dispositif observé. Si la collecte est poussée (push), l'analyse de flux appliquée à la collection renseigne sur l'activité perçue[252] de ce dispositif. La validité des connaissances produites repose alors sur la qualité d'enveloppement du dispositif d'observation, celui-ci devant coller en continu au dispositif observé sans en perturber le fonctionnement. Si le flux est de ce point de vue valide et s'il peut être considéré comme isomorphe avec ce qu'il traduit, alors il permet une analyse relative à ce flux des dynamiques de l'activité effective. Conjointement, l'existence de motifs structurels renseigne sur les configurations et permet une analyse des mises en relations ainsi identifiées. Dans la complémentarité de ces deux analyses, c'est une démarche systémique que nous privilégions. Par la suite, l'authentification des acteurs lorsqu'elle est possible, orientera vers une analyse individualisée des contributions. C'est à partir de cette analyse que des phénomènes et des comportements remarquables (récurrents, atypiques, etc.) pourront être recherchés comme autant d'axes analytiques.

2.3. L'incidence des *Big Data* dans la production de connaissances

Le relâchement des exigences formelles établies sur la qualité des données à l'échelle des *Big Data* a une incidence sur les mécanismes et la nature des connaissances produites. Parmi les contraintes relâchées nous avons évoqué dans le chapitre précédent la *complétude*, *l'exhaustivité* ou la *représentativité*. Les considérations techniques et méthodologiques sur la collecte nous situant à l'échelon de la ressource, nous avons principalement envisagé ces contraintes appliquées sur les données. Dans ce contexte, ces exigences portent sur la capacité de passer d'un formalisme numérique, celui de la ressource, à un autre[253], celui de la collection. Le relâchement de contraintes accompagnant le

[251] Littéralement *un retour sur trace* ce qui prend tout son sens ici.

[252] Tenant compte des limitations expérimentales d'accès (flux contrôlé) ou de capture (bande passante, etc.).

[253] Il s'agit plus précisément d'une *transduction* au sens de la théorie des langages, c'est-à-dire comportant des transformations complexes et non pas une traduction terme à terme.

changement d'échelle s'est traduit par l'adoption de modèles de représentations de données alternatifs, moins rigides[254] et de modalités inférentielles inductives adaptées.

2.3.1. Représentation des données et appropriations heuristiques

Sans pour autant sonner le glas des modèles et méthodes associés au type relationnel incarné par le format SQL[255], l'évolution des modèles de représentation, impulsée par la prise en compte des *Big Data* a, sans conteste, transformé le rapport aux données numériques.

D'un point de vue pratique, l'extension de la typologie des modèles permet de choisir un format spécifique en fonction de la nature des données et des traitements envisagés. Cette spécialisation répond à des attentes diverses, qui peuvent être :

- de nature technique afin de réduire les coûts de stockage, d'accès aux données et plus généralement d'améliorer la performance relative aux traitements spécifiques ;
- de nature formelle afin de faciliter la modélisation ou la production de requêtes tout en améliorant l'intelligibilité des données dans leur contexte d'usage.

Nous insistons plus particulièrement sur ce dernier point. L'abstraction proposée dans ces modèles masque une certaine complexité technique et concourt à rendre moins hermétique les langages de manipulation et les formats de représentation des données. Cette évolution favorise une prise en charge plus précoce des données par des chercheurs non informaticiens. Dans le contexte de collections, cela passe par une connaissance préalable des choix représentationnels. La mise en place des plans de gestion de données (DMP) est une proposition documentaire normative qui va dans ce sens. Mais ce ne sont que des moyens. La prise en charge ne peut être effective sans acculturation aux formalismes des données numériques et sans médiations appropriées.

La simplicité apparente des outils graphiques qui se développent en interface des systèmes représentationnels constitue une avancée certaine dans ce sens. Mais elle renferme aussi un leurre qui est celui d'une lecture immédiate et intuitive des données. Le développement des techniques de visualisation favorise la prise d'informations sur les données et sur les résultats de calculs qu'ils projettent dans un espace graphique. Les vues ainsi réalisées sont indéniablement heuristiques mais il s'agit d'abord de représentations produites par des matériels et d'un empilement de modèles qui s'interposent entre le lecteur et les données. Le travail de décodage qui est demandé impose à l'analyste une connaissance fine du langage graphique et des modèles sous-jacents de traitements.

De nombreux articles, notamment dans les SIC, mettent en garde vis-à-vis des dérives méthodologiques possibles que ces outils et méthodes peuvent introduire. Nous ne pouvons que souscrire à la mise en garde sans pour autant suivre la critique jusqu'à son terme qui est, dans de nombreux cas, un rejet strict de l'approche numérique et expérimentale.

---

[254] Modèles alternatifs, identifiés comme NoSQL.

[255] Qui peut d'ailleurs être utilisé (ou non) dans les couches représentationnelles profondes des modèles réalisant une abstraction plus adaptée. Par ailleurs des évolutions telles que PostgreSQL poursuivent l'adaptation des modèles relationnels au *Big Data*.

2.3.2. Un renouveau dans l'approche des données et des modèles

Plus avant, aborder la production de connaissances dans toute sa généralité impose de dépasser la limite formelle de la donnée numérique pour l'aborder du point de vue informationnel. Le but est de ne pas s'arrêter à l'information mise en forme et diffusée numériquement (Web et API) mais de s'intéresser aux processus de production et de sélection qui s'opèrent en amont (zone grisée de la figure Fig40). Le processus d'éditorialisation qui contrôle cette sélection d'informations puis sa mise en forme numérique publicisée fait partie des questions de recherche permettant d'aborder l'*informationnalisation* et de là le procès de communication.

Expliciter ce processus est l'un des objectifs de l'observation expérimentale des pratiques info-communicationnelles. La compréhension des processus précédents, au travers de l'analyse des données, est une question ouverte à laquelle il n'est pas aisé de répondre en dehors de cas particuliers. En revanche, l'articulation entre information reconstruite[256] (seule accessible) et représentation numérique disponible, permet d'étendre le questionnement sur la production de connaissances.

Les notions d'exhaustivité[257] ou de représentativité interviennent aussi dans ce questionnement. Ces notions, dont l'origine est méthodologique, sont alors investies d'une valeur épistémologique liée aux conditions normalisées de la production scientifique. Avant d'être largement mobilisés par les modèles formels des bases de données, les critères qualitatifs que ces notions expriment ont été associés aux modèles d'analyse statistique adoptés par les Sciences humaines et sociales au début du XIXe siècle (Boullier, 2015, p5). Ces critères s'appliquent à l'échantillonnage d'une population statistique dont il s'agit de conserver les caractéristiques intrinsèques pertinentes dans l'échantillon. La nécessité d'une réduction contrôlée de la population, conforme aux modèles statistiques, légitime de tels critères. L'apport indéniable de ce « *contournement…[par les statistiques]…a fourni aux Sciences sociales un raccourci bien pratique* » pour reprendre les propos de Tommaso Venturini et Bruno Latour citant les travaux d'Alain Desrosières (Desrosières, 1993), (Venturini, Latour, p3). Peut-on pour autant considérer qu'en dehors de ces modèles, ces critères n'ont pas de sens ? Répondre oui serait s'en tenir à une raison de pure forme. La nature intuitive de ces notions ne les écarte pas définitivement d'une démarche empiriste ou constructiviste. L'incrémentation des collections est d'ailleurs souvent dirigée par ces critères. Il ne s'agit donc pas de les rejeter mais d'en avoir une lecture différentiée, relative aux choix méthodologiques retenus. Plus généralement, les méthodes par échantillonnage ne sont pas envisagées dans une approche *Big Data* des collections parce que la réduction n'est pas nécessaire et qu'elle est même contre-productive vis-à-vis des techniques des *Big Data*.

Cela ne signifie pas non plus qu'il faille rejeter les méthodes par échantillon. L'exploration de collections massives peut amener à des réductions pour lesquelles l'échantillonnage redevient pertinent par un retour aux conditions classique de l'analyse. Globalement, la difficulté à laquelle

---

[256] Peut-être pourrait-on parler de ré-informationnalisation comme d'un processus dual dans lequel l'analyste est engagé ?

[257] La complétude est un terme informatique qui n'a pas lieu d'être à ce niveau.

nous sommes confrontés réside dans l'articulation des méthodes analytiques ; la question ouverte est par nature méthodologique.

3. Perspectives et enjeux de l'analyse des traces numériques

L'intégration environnementale de plus en plus poussée des fonctions info-communicationnelles nous incite à franchir l'étape consistant à aborder non plus les machines ou les objets identifiés à des technologies mais des dispositifs complexes. Ces dispositifs se caractérisent par leur omniprésence, spatiale et temporelle, suivant des configurations circonstancielles variables. Dans ce contexte, les technologies n'occupent plus le premier plan, ce qui remet en question l'ensemble des articulations, y compris symboliques, dont elles sont investies. La complexité de l'usage se déporte, passant des artefacts spécialisés et circonscrits, à un environnement continu de potentialités (*affordances*[258]) info-communicationnelles. Parallèlement, la généralisation de la traçabilité comme principe, et du traçage comme méthode de représentation personnalisée, conditionne durablement les logiques d'offre et d'organisation de services numériques.

C'est selon cette double perspective que nous proposons d'aborder les enjeux scientifiques de l'analyse des traces numériques pour les disciplines des Sciences humaines et sociales, et plus particulièrement pour les Sciences de l'information et de la communication. Nous aborderons à la suite mon positionnement vis-à-vis de ces enjeux dans l'orientation de mes travaux et projets de recherche.

3.1. Enjeux scientifiques et disciplinaires

Dans ce mémoire, nous avons pris soin de distinguer plusieurs niveaux de représentations numériques correspondant aux transformations successives apportées aux traces numériques issues de l'observation pour constituer *in fine* des jeux de données (corpus) pour l'analyse. Ces transformations déterminent un processus d'élaboration incrémental des collections ; nous avons abordé ce processus d'un point de vue méthodologique et nous l'avons justifié dans un contexte de production et de communication scientifique. La collection résultante est un objet représentationnel sous déterminé[259] permettant de produire un ensemble de corpus de données en réponse à une classe de problématiques analytiques.

La dimension méthodologique qui nous a guidé fait apparaitre des enjeux scientifiques propres à la démarche de collecte dans un contexte d'observation. Le questionnement qualitatif portant à la fois sur le processus de collection et sur son résultat est un aspect important et délicat du travail empirique que nous conduisons. L'accumulation d'expériences apporte une base réflexive à ce sujet. Cela se traduit en partie par un ensemble de bonnes pratiques que nous avons présentées. Il est cependant difficile d'aller au-delà de quelques principes méthodologiques, notamment au regard de questions comme par exemple, la maximalisation de la classe des corpus que la collection permet

[258] Au sens de James. J. Gibson.

[259] C'est-à-dire non entièrement spécifié et finalisé.

d'engendrer pour un dispositif donné. Dans le prolongement de ces enjeux scientifiques, d'autres plus généraux et de nature épistémologique apparaissent dès lors que l'on interroge la nature et l'étendue de la classe des dispositifs info-communicationnels (observables) supportés dans un contexte numérique. Le cadre d'observabilité que nous avons envisagé est celui de la production de traces associées à la publication de contenus ou à la publicisation de données. Outre les limitations (légales, techniques, etc.) qui encadrent le processus d'observation, il n'est pas possible à l'heure actuelle de répondre à un questionnement théorique sur la portée heuristique des collections optimisées que l'on peut produire.

Résoudre cette question supposerait, en premier lieu, une capacité à se situer dans l'espace des connaissances produites, inférées à partir de caractérisations de données. Unifier les espaces de données et ceux des connaissances est un sujet ouvert depuis les débuts de l'informatique et qui est loin d'être épuisé. En second lieu, cela supposerait que l'on soit en mesure d'établir un cadre général de recherche sur les dispositifs info-communicationnels. Tracer les contours de ce cadre, signifierait que l'on soit d'une part, en mesure de formaliser et décrire cette classe de dispositifs et d'autre part, que l'espace problématique soit clairement posé.

Or il apparait que ce dernier, de par sa complexité, ne s'investit que très progressivement par tâtonnements successifs. De notre expérience, il ressort que bien souvent on n'est pas en mesure de répondre à une question théorique à partir d'une collection pourtant réalisée à cet effet. En revanche, la disponibilité des données de collection occasionne des allers-retours entre hypothèses et résultats d'analyses. Fréquemment, des résultats divergents ou décevants amènent à reformuler ou affiner un questionnement. Parfois même, la reformulation d'un questionnement peut être considérée comme un résultat en soi. En ce sens, les collections ont un intérêt heuristique évident. Il arrive par ailleurs, que l'on ne puisse rien inférer à partir des données en dehors de cas particuliers que peuvent néanmoins étayer des intuitions. Le matériau de la donnée, notamment issu d'énoncés linguistiques, peut avoir une complexité qu'il n'est pas toujours possible de manipuler en dehors du contexte de l'énonciation.

Au niveau macro, la notion de dispositif comporte une plasticité et une dynamique influencée par l'évolution de l'offre technologique et des logiques sociales d'usage. Délimiter une classe de dispositifs opérante d'un point de vue scientifique revient à dépasser ces variations pour identifier les archétypes représentatifs de son extension. Deux axes se dessinent : l'un porte sur l'étude structurelle des dispositifs, des agencements de ressources, des rôles des acteurs, etc. ; l'autre porte sur l'étude des processus info-communicationnels et des médiations organisées par les dispositifs. Ces deux axes complémentaires ont une place historique dans le champ des SIC.

Il ne nous semble pas que le changement de paradigme que traduit la notion de dispositif info-communicationnel bouleverse radicalement le champ disciplinaire des SIC. Cette perspective poursuit naturellement l'élargissement déjà engagé avec l'approche située des phénomènes et des pratiques info-communicationnels.

À l'issue de nos travaux, l'évolution qui nous parait la plus remarquable est celle qui se traduit du point de vue des dispositifs numériques, dans la mesure où le processus de *datafication* occupe

désormais le premier plan. Nous en avons souligné l'importance sur le plan économique mais aussi sur le plan politique.

Ce processus appelle du point de vue des SIC un renforcement de l'analyse critique ainsi qu'un positionnement éthique et moral au regard de l'évolution de la société que la production de données massives et leur interprétation rapide influencent. Si comme nous le pensons, le processus de *datafication* est étroitement lié à celui d'*informationnalisation* porté par les dispositifs info-communicationnels et donc médiatiques, alors l'engagement des SIC sur ces enjeux n'en est que plus important. L'approche que nous poursuivons sur l'observation instrumentale ne nous écarte pas fondamentalement de ces débats. Notre responsabilité scientifique est engagée mais elle ne peut pas se traduire par la définition d'une zone d'exclusion dans nos pratiques. Nous devons au contraire, poursuivre nos investigations sur la *datafication* à la fois du point de vue des dispositifs observés et de celui de l'observation.

L'éditorialisation des traces numériques disponibles aux bornes des API comme ressource analytique pour les Sciences sociales soulève un questionnement épistémologique que nous ne pouvons pas éviter. L'absence de neutralité de la trace numérique d'usage nous permet-elle de fonder une connaissance pertinente et scientifique propre à soutenir des modèles et des théories ?

Notre pratique du Web nous amène à considérer la fonction de réduction comme une conséquence majeure de l'éditorialisation. En effet, comme nous l'avons déjà souligné, les plateformes de services doivent être visibles. La publication sur le web assure une partie de cette visibilité stratégique que vient compléter la disponibilité de données de traces. Ces dernières alimentent une *externalisation ouverte* ou *crowdsourcing* qui contribue au rayonnement de l'entreprise et potentiellement à son futur développement. Dans ces conditions, la nature des altérations produites dans l'éditorialisation doit être suffisamment contrôlée pour ne pas tout dévoiler des données internes sans trahir leur potentiel heuristique.

La réduction opérée ne permet pas de raisonner dans la généralité des données complètes enregistrées dans le système d'information des plateformes. Dans l'hypothèse où la plateforme de services ne manipule pas intentionnellement l'information qu'elle produit à l'occasion de l'éditorialisation, la réduction maintient une fonction analogique entre les deux types de données originelles et extraites. Les résultats obtenus sont moins complets et moins précis, mais ils conservent les tendances analogues à une analyse équivalente qui aurait été opérée sur les données originelles. Mais cette hypothèse reposant sur la confiance dans les logiques de la publicisation est par nature fragile.

La situation à laquelle nous sommes confronté n'est cependant pas très nouvelle. L'échantillonnage systématisé dans les méthodes des Sciences humaines et sociales comporte lui aussi des limites que nous avons tendance à normaliser. La différence, mais elle est de taille, est que, dans ce cas, c'est le chercheur qui opère la réduction, il en est donc responsable. Dans le cas des traces numériques d'usage, nous ne maitrisons pas les conditions initiales de leur production. Cependant, cette limitation a un effet favorable aux pratiques de la recherche puisque les conditions d'analyse sont de fait normalisées.

3.1.1. Nouvelles approches et transdisciplinarité

La prise en compte de la disponibilité massive des données associées notamment au Web en tant qu'espace de leur production et de leur circulation fait apparaitre des courants potentiellement structurants et susceptibles d'atteindre notre discipline. Les mouvements issus de *recherche numérique*[260] comme les déclinaisons des *Internet Studies*[261] que sont les *Cultural Analytics*[262], les *Digital Methods*[263] figurent parmi ces émergences qui traversent les disciplines des Sciences humaines et sociales et questionnent leurs pratiques.

Dans l'analyse qu'ils font de la «*redistribution des méthodes*», Jean-Claude Plantin et Laurence Monnoyer-Smith associent à ces deux propositions précédentes, celle plus ancienne des humanités numériques[264] (*Digital Humanities*) (Plantin, Monnoyer-Smith, 2011, 2013, p44). Les humanités numériques se sont constituées et institutionnalisées à partir de projets de numérisation (*digitization*) et d'analyses numériques de documents propres aux besoins scientifiques des disciplines des humanités. Par nécessité, ce rapport instrumental au numérique est encore très marqué. Il diffère en cela des deux autres approches qui visent aussi (voire exclusivement pour les *Digital Methods*) l'exploration de ressources numériques natives (Rogers, 2009, p5). Le terme *natif* est ici employé dans le sens qu'on lui donne en informatique. Il caractérise les données envisagées dans les formats et les structures de données utilisés dans les traitements qui les ont produites. Cette définition permet de distinguer ces données de celles issues d'une numérisation. La différence est soulignée en anglais dans l'usage du terme *digitalization* pour caractériser ces ressources digitales naturelles.

L'écart entre les deux perspectives précédentes est cependant appelé à se combler pour plusieurs raisons. Comme le suggère Imad Saleh et Hakim Hachour, la catégorie des humanités est désormais considérée plus largement comme celle des Sciences de l'homme et de la société (Saleh, Hachour, 2012, p4). Ensuite, les différents manifestes pour les humanités numériques produits au tournant des années 2010 (séminaire UCLA Mellon mai 2009[265], ThatCamp Paris 2010[266]) ont intégré le Web dans l'espace documentaire ainsi que les transformations du Web collaboratif dans leurs pratiques (Presner, 2010). Enfin, les problématiques numériques actuelles comme, l'archivage de collections

---

[260] Nous adoptons l'expression générique telle qu'elle est définie par Jean-Claude Plantin et Laurence Monnoyer-Smith (Plantin, Monnoyer-smith, 2013, p42).

[261] Considérées dans leur troisième période selon Barry Wellman, c'est-à-dire depuis les années 2000 (Wellman, 2004). Pour être plus précis, à partir du changement qui consiste à ne plus considérer le Web comme un espace de virtualités mais comme un espace en prise directe avec le réel (Jones, 1999).

[262] Désignation proposée en 2005 par Lev Manovich (Manovich, 2016) qui n'a pas de traduction française attestée

[263] Dénomination générique mais faisant écho dans le contexte présent aux travaux de Richard Rogers (Rogers, 2009). Il n'a pas non plus d'équivalent français attesté.

[264] Lou Burnard fait état des travaux précurseurs de Roberto Busa (Corpus Thomisticum) à la fin des années 1940 comme point de départ des humanités numériques (Burnard, 2012, p48)

[265] http://manifesto.humanities.ucla.edu/

[266] http://tcp.hypotheses.org/318

numériques, la mise à disposition des corpus et des données de la recherche[267] (*open data*) ainsi que la facette documentaire qui les accompagne sont largement portées par les humanités numériques.

Pour les SIC, les problématiques documentaires et représentationnelles soulevées par les humanités numériques ont favorisé leur ancrage au sein de la discipline dès l'origine. Cette légitimité se maintient depuis, renforcée par les évolutions qui affectent réciproquement les humanités numériques dans leurs problématiques et la discipline dans son épistémologie. Le dossier numéro 8[268] « *humanités numériques et SIC* » de la revue française des Sciences de l'Information Communication atteste de cette proximité. Bien que ce dossier n'épuise pas le sujet, on peut constater à sa lecture que les humanités numériques sont envisagées dans une ouverture large et englobante de la recherche numérique en SHS au point de se confondre avec elles. On retrouve néanmoins une originalité de l'approche des SIC dans l'élaboration d'une pensée critique qui interroge « *les concepts et le statut contemporain de l'humanisme et son épistémè* » (Cormerais & Al., 2016, p4).

Comme leur nom l'indique, les *Cultural Analytics*[269] contiennent une réflexion méthodologique sur l'analyse culturelle au sens large de phénomènes associés aux enregistrements numériques. Ce mouvement initié au début des années 2000 par Lev Manovich, conduit ce dernier à s'interroger sur les analyses permises dans le cas de données numériques massives et tout particulièrement sur les possibilités offertes par la visualisation. Le contexte des données naturelles (massives, dispersées, fragmentées et faiblement structurées) ainsi que la complexité des analyses sur les objets numériques qui sont conjointement visés dans ce programme d'études le situent d'emblée dans les *Big Data*. Le recours à des traitements informatiques adaptés en conséquence (s'appuyant sur une infrastructure de type haute performance) est de ce fait considéré comme incontournable (Manovich, 2009, p7). S'il privilégie les données culturelles (*cultural data*) associées à des objets numériques bien identifiés (photo, vidéo, etc.) et à des gisements structurés (Flickr, Youtube, etc.), les travaux engagés dans ce mouvement sont transposables à d'autres contextes de données moins qualifiées. L'évolution des *Cultural Analytics* conduit ces promoteurs à un rapprochement avec les *Digital Humanities* et le *Social Computing* (Manovich, 2016). Le sens de ce rapprochement avec les humanités numériques est la confrontation de ces différentes approches[270] sur les objets culturels dans le but de dégager une réflexion historique et critique et de croiser les expériences afin d'ouvrir de nouvelles voies. Le rapprochement simultané avec le *Social Computing* conforte l'engagement technologique tout en intégrant la dimension collaborative des formes de production.

À la différence des *Cultural Analytics*, les *Digital Methods* portent exclusivement sur des données natives (*born-digital data*). Le programme tracé par Richard Rogers est d'ordre méthodologique (Rogers, 2009). L'hypothèse fondamentale justifiant une démarche spécifique est de nature situationniste. Pour Rogers en effet, la compréhension du contexte et des mécanismes de

---

[267] Voir §3.1.2.

[268] https://rfsic.revues.org/1778

[269] Le site http://culturalanalytics.org/ est une vitrine du mouvement engagé par Lev Manovich.

[270] Y compris la différence de nature des données *digitized* vs *digitalized*.

production des données fait partie de leur analyse. Comme il l'explique, on ne peut aborder les collections de données sans évoquer les discontinuités ou l'instabilité du service durant la collecte, et plus généralement les termes et conditions de leur publicisation (Rogers, 2016, pp.9-10). Ce point de vue est analogue à ce que nous défendons dans ce mémoire. Son propos va au-delà des conditions de publicisation puisqu'il évoque la nécessité d'une connaissance fine des algorithmes justifiant la production de ces données dans le service (*Ibid.*, p2).

L'autre caractéristique de la démarche consiste à rechercher les spécificités techniques que les données natives impliquent plutôt que de chercher à transposer des méthodes classiques ou associées à des données numérisées (*digitized*). L'approche se distingue ainsi du courant des *virtual methods* qui promeut la transposition des méthodes qualitatives des SHS dans l'espace numérique (Hine, 2000, 2005 citée par Marres 2012). Dans l'esprit, nos travaux sont proches des *Digital Methods*.

Les *Cultural Analytics* ou les *Digital Methods* se positionnent comme des terrains d'expérimentations ciblés en dehors de cadres méthodologiques et épistémologiques fixés dans les disciplines. Ces mouvements transdisciplinaires par nature, bien identifiés par leur objet, parfaitement outillés dans leur communication, portés par des structures institutionnelles et des logiques de fonctionnement proches de structures de projets, répondent quasiment à un cahier des charges d'innovation. Si la forme peut interroger, il est indéniable que de tels mouvements qui peuvent n'être que transitoires, alimentent les débats scientifiques et disciplinaires par leurs résultats et le questionnement qu'ils engendrent.

L'intérêt suscité par ces deux mouvements dans les SIC est encore peu perceptible en tant que proposition scientifique, bien que les travaux de Lev Manovich ou de Richard Rogers soient assez bien référencés dans les publications. L'article de Jean-Christophe Plantin et de Laurence Monnoyer-Smith examine les apports possibles pour les champs d'études des SIC de ces deux voies. Outre l'intérêt méthodologique et heuristique souligné par ces auteurs, l'apport principal de leur article reprend les travaux de Noorje Marres (Marres, 2012) sur les « *effets de redistribution des méthodes* »[271] qui ont lieu dans la recherche numérique (Plantin, Monnoyer-Smith, 2013, pp.44-45, pp.60-61). L'hypothèse est inspirée des travaux de Bruno Latour en sociologie des Sciences. Elle concerne le traçage de la redistribution des méthodes comme moyen d'information sur l'évolution des pratiques et des interactions sociales à l'œuvre entre les acteurs de la recherche académique, industrielle ou de la société (Marres, 2012, p145). Le point de vue défendu est celui d'une redistribution plus fine et significative que l'affirmation clivée du transfert (académique vers industrie) ou du monopole industriel de la recherche numérique en SHS (Savage, Burrough, 2007). Dans l'analyse de Noorje Marres, la redistribution est croisée. Elle passe par l'implémentation des méthodes dans les plateformes de services[272] et la réappropriation des données de services dans les

[271] La redistribution signifie en l'occurrence qu'on ne peut pas cloisonner ni enfermer la recherche numérique dans un seul domaine, une institution, un type d'activité, etc. mais que l'on doit la considérer au travers d'un espace de mobilisation et de recomposition d'une pluralité d'acteurs (Marres, 2012, p140).

[272] Elle évoque par exemple les co-citations, etc.

méthodes des SHS. Cette mise en regard, n'est pas sans évoquer les hypothèses de la sociologie de l'usage. Jean-Claude Plantin et Laurence Monnoyer-Smith suggèrent de prolonger cette démarche.

Nous partageons ce point de vue qui permet de poser un cadre propre à définir une *tracéologie numérique* adaptée au contexte d'étude des traces numériques d'usage. Un tel chantier visant à stabiliser les méthodes et éprouver les hypothèses nous parait indispensable. Mais cette voie ne doit pas nous rendre obtus. D'autres voies doivent être envisagées en parallèle, comme celles visant à obtenir une plus grande ouverture de l'espace des données du Web.

3.1.2. Données publiques et espace de publication des données

La prise en compte de la publicisation des données numériques est un vaste débat dont les dimensions économiques et politiques sont devenues très marquées et difficilement décomposables.

L'expérience de l'industrialisation des méthodes numériques nous a permis d'approcher la dimension économique principalement sous l'angle de la chaîne de la valeur, comme nous en témoignons à plusieurs reprises dans ce mémoire. Lorsqu'il s'agit d'établir des profils personnels, les limites de la caractérisation endogène, c'est-à-dire spécifique à un dispositif sont apparues nettement. Rechercher des informations en dehors de ce périmètre et constituer des données exogènes est alors un enjeu déterminant. Cet enjeu peut être abordé de deux façons, soit en adoptant le point de vue du marché des données, soit en considérant qu'il existe une ressource ouverte commune et publique : le Web. Ce dilemme rencontré dans une start-up est analogue à celui que nous rencontrons à l'occasion de projet de recherches dans le domaine numérique. La concentration actuelle du marché numérique sur un petit nombre d'acteurs verrouille la possibilité d'entreprendre dans les deux cas. Le ticket d'entrée est élevé d'autant qu'il s'agit, comme dans une partie de poker, de miser pour voir.

Le Web joue actuellement différentes fonctions l'assimilant à un espace public sans que les règlementations nationales et transnationales permettent de le considérer comme tel. Ainsi, l'énorme volume de publications, accessible à tous, ne peut pas être exploité sans se trouver dans l'illégalité que l'*exception recherche*[273] ne couvre que partiellement. La question politique a bien été posée, réduisant un peu l'écart existant entre les politiques européennes et celles des pays anglo-saxons.

La mise en collection des données, que nous associons à l'observation des dispositifs info-communicationnels porte un double projet. Un projet éditorial de publication de travaux de recherche et un projet collaboratif de publicisation de données collectées. En l'état actuel de la loi, ce partage communautaire nécessité par l'évolution des pratiques de la recherche ne peut pas être envisagé sans une stérilisation importante des données. Les contraintes liées aux données nominatives n'en sont pas la principale cause. Ce sont plutôt les craintes associées à la

[273] Amendement n°180 à l'article 18 du code de la propriété intellectuelle modifié à l'occasion de la loi *pour une République numérique* (publiée au JO le 8 octobre 2016). http://www.assemblee-nationale.fr/14/amendements/3399/AN/180.asp. Cet amendement prend en compte la fouille des contenus du Web (Text and Data Mining).

représentation personnalisée qui agissent comme un frein. Parmi celles-ci, il y a la possibilité de recoupement que permet la multiplication des tables de données individuelles. À multiplier les descriptions individuelles, l'identité se précise par des voies détournées. L'arrivée de référentiels normatifs et la sémantisation du Web participent de cet effet lié à la densification informationnelle, de manière inévitable. Éliminer le risque qu'une référence individuelle ne devienne explicite (personnelle) passe par l'élimination de toutes les références périphériques y compris catégorielles, telles que les lieux, les institutions, les événements, etc. auxquels l'individu est associé dans les représentations numériques. Procéder de la sorte, ne peut que dénaturer les enregistrements et vider de leur sens les collections.

Conjointement, la mise en œuvre de catégories analytiques de caractérisation sociologique (profilage) nécessaire au raisonnement scientifique, et classique parmi les méthodes des SHS, apparait suspecte dès lors que les catégories sont numériques. Attribuer une orientation politique, religieuse ou sexuelle afin de constituer des classes de profils pour l'analyse ne peut pas s'envisager à la légère. La manipulation d'enregistrements numériques est immédiatement assimilée à du fichage.

Dans le même sens, l'exploitation des métadonnées provenant de l'archivage institutionnel réalisé par la Bibliothèque Nationale de France (BNF) est une avancée enthousiasmante mais toujours pauvre par rapport aux contenus numériques effectivement archivés qui demeurent inaccessibles. Dans une économie numérique et concentrée, l'échelle transnationale s'impose dans le débat sur la patrimonialisation du Web et de ses évolutions. La mission d'archivage institutionnel de la BNF, entérinée par la loi de 2006, se résume en l'état au domaine France (.fr) alors que les ressources documentaires du Web n'ont pas de frontières. De très nombreuses ressources, comme les réseaux sociaux, échappent de ce fait à la mission de la BNF alors que dans le même temps, Twitter reverse ses bases de données à la bibliothèque nationale du Congrès.

Au travers de ces exemples qui font notre quotidien, nous ne pouvons que constater l'état de transition numérique dans lequel nous sommes impliqués en tant que chercheurs et citoyens. L'exploitation de données ouvertes n'est encore qu'un vœu que nous formulons. Des aménagements sont encore nécessaires, n'appelant pas systématiquement un appareillage législatif. Ils peuvent s'établir selon un principe analogue au *fair use* appliqué dans les pays anglo-saxons. Cette liberté de pratiques dans l'espace de la recherche est indispensable pour maintenir la possibilité d'un débat ouvert dans une société dont le devenir est présenté comme celui de la connaissance. L'horizon 2020 fixe une nouvelle étape après celle de 2010 dans la construction d'un espace européen de la recherche.

3.2. Enjeux et perspectives personnels

Le mémoire que je présente suit une trame historique dont le fil directeur est celui de l'évolution des pratiques instrumentées soutenues par des technologies numériques. L'origine de ce fil correspond à la naissance du Web dans le monde académique alors que dans le même temps se consolidaient les usages de la télématique. À l'autre extrémité de ce fil, nous assistons au double phénomène :

- la dilution progressive de la composante technologique dans l'espace physique ne maintenant qu'une abstraction de services ;
- l'incorporation de pratiques techniques permettant de composer avec l'espace physique pour conduire une activité.

Tout au long de ce fil, ma démarche de recherche a été portée par la volonté d'être au plus près de ces transformations pour comprendre leurs logiques et leurs portées dans l'espace social. L'observation s'est très vite imposée comme un moyen et sa difficulté comme un challenge. Les projets auxquels j'ai participé ont été conduits pour la plupart dans l'objectif de défricher un champ d'étude ou d'innover dans la manière de l'aborder. Cette position d'innovation dans la pratique de la recherche m'a sensibilisé à la question méthodologique. J'ai également pu constater que les outils de l'observation peuvent être le soutien d'une formalisation computationnelle de phénomènes observés. Ainsi, l'observation numérique devient elle-même objet de recherche. C'est dans ce sens que je situe mon intérêt pour les *Digital Methods*, ou les *Cultural Analytics* dont je trouve les apports particulièrement stimulants, y compris d'un point de vue épistémologique. C'est au cours de cette année pendant laquelle je me suis investi dans l'écriture que j'ai pleinement envisagé l'intérêt du renouvellement des humanités numériques pour la démarche scientifique en SIC. La parution du tout récent numéro 10 de la *Revue Française des Sciences de l'Information et de la Communication* (RFSIC)[274] a été l'occasion d'affirmer cet engagement au sein d'un groupe de travail portant sur l'inscription des *Digital Studies* dans le champ disciplinaire (Ref.44).

Depuis 2010, mes travaux suivent le double mouvement de la convergence médiatique et de l'affirmation du Web, étendu aux réseaux sociaux, en tant qu'extension numérique de l'espace public. Cette orientation scientifique personnelle me parait durable parce qu'elle ouvre sur des champs de recherches interdisciplinaires féconds et bien positionnés dans le contexte académique grenoblois. Elle constitue aussi un choix éthique qui consiste à faire exister dans le cadre de son activité des convictions et des valeurs personnelles. L'existence d'un espace numérique considéré comme lieu d'expression public libre, mais aussi comme bien commun me parait être un enjeu qui à mes yeux justifie pleinement cette revendication militante. Ce positionnement scientifique et citoyen est en lien avec le projet de formation aux *métiers du livre et de l'édition* dans lequel je suis très investi et dont l'assise scientifique doit être renforcée.

3.2.1. Projets scientifiques en cours

L'analyse des données de carrières extraites de LinkedIn se poursuit dans le contexte des métiers du journalisme suivant une perspective de collaborations internationales (Brésil, États-Unis, Angleterre). La plateforme MEDIASWELL est proposée comme support commun à ces collaborations. Outre l'intérêt du projet scientifique, ce programme est une opportunité pour la mise en œuvre d'un portage expérimental en dehors du cadre de sa fabrication. À cette occasion, l'objectif technique est de finaliser l'ouverture des codes sources et de les diffuser dans le contexte

[274] https://rfsic.revues.org/2543

du logiciel libre. Franchir cette étape constitue un aboutissement favorisant la dissémination des méthodes et la possibilité d'étendre les fonctionnalités de la plateforme dans un processus d'élaboration collaborative ou *crowdsourcing*.

L'évolution de la plateforme MEDIASWELL se poursuit en parallèle de la mise en place d'un cadre d'analyse empirique pour l'ANR RSJ-MéDiS[275] (*Responsabilité Sociale des Journalistes, Médias Diversité Sport*) afin de supporter l'analyse des micro-événements médiatiques que sont les retransmissions télévisuelles. L'objectif est d'établir des collections dont la trame temporelle associe les différentes réactions engendrées par les commentaires sportifs produits en direct ou disponibles à la demande sur YouTube. Il s'agit d'articuler des espaces connexes de production discursives : la retranscription des commentateurs, les captations sur Twitter et les commentaires associés aux pastilles vidéo sur YouTube. Les collections sont finalisées afin de permettre l'analyse temporelle des contenus de publications.

En complément des travaux engagés dans le programme RSJ-MéDiS sur la question de l'expression de la diversité, Une collection de longue durée (8 mois) est en cours de réalisation dans le cadre des programmes récurrents portant sur les campagnes électorales. Elle est mise en œuvre au sein du laboratoire PACTE en relation avec des travaux engagés dans les laboratoires de Sciences politiques de Paris et de Bordeaux. Dans le cas présent (échéance 2017), ce sont les questions connexes au programme RSJ-MéDiS, associées aux enjeux politiques de la *diversité* et de *l'identité* qui sont envisagées. Dans le contexte sociétal actuel, nous faisons l'hypothèse que ces notions, en lien avec la définition de la démocratie, articulent des engagements militants et des discours politiques. De cette manière nous pourrons analyser de manière contrastée la nature et les contenus des discours en lien avec ces notions dans les deux contextes différents.

3.2.2. Projets scientifiques à venir

L'augmentation des volumes de données comme la complexité des représentations numériques supposent, en amont de toutes considérations méthodologiques, d'aligner l'environnement informatique de la recherche en Sciences humaines et sociales sur des standards technologiques du *Big Data*, voire du calcul intensif (HPC). Réunir les conditions

Pour prolonger et soutenir les actions de recherche engagées, je me suis investi dans deux projets.

- Le premier projet, *READ / ERDA EBooks Recommendations and disengageable algorithms,* concerne l'évolution des pratiques sociales sur le Web qui renouvelle les logiques et les mécanismes de la diffusion du livre numérisé et numérique. Ce projet repose sur la collaboration engagée à l'occasion de l'ARC 6[276], entre les laboratoires ELICO (EA 4147) et PACTE (UMR 5217) associant les laboratoires du GRESEC (EA 608) et du LIG (UMR

[275] ANR 15-CE26-0006-01. Collaboration organisée entre les laboratoires : PACTE (UMR 5194), CRAPE (UMR 6051), Praxiling (UMR 5267 CNRS), Geriico (EA 4073) et URePSSS (EA 4110).

[276] Action de Recherche Régionale pilotée par Françoise Paquienséguy du laboratoire ELICO. Ce projet fait l'objet d'un dépôt de projet UGA (*TidEBook*) dans le cadre IDEX - Initiative Recherche Stratégique (IRS). Par ailleurs, une demande de création de Groupe d'Étude et de Recherche (GER) est également engagée auprès de la SFSIC.

5217). Mon investissement dans ce projet est essentiel, compte tenu des enjeux d'articulation formation et recherche. Ce cadre de projet est aussi l'occasion pour moi d'aborder la question des collections de données structurelles et temporelles du Web associées à une problématique de cartographie dynamique des ressources participant à la diffusion et à la distribution, légale ou non, du livre numérique.

- Le second projet, *HOPISICITE Hospitalité – sociétés civiles européennes – émotions,* interroge les façons dont les émotions circulant dans les médias influencent la mobilisation dans les sociétés civiles européennes, notamment dans le contexte de la "crise des réfugiés". La question de l'expression des émotions dans l'espace public poursuit un travail que j'avais précédemment engagé dans le contexte du projet PEPS HuMaIn[277] pour le contexte de l'élection européenne de 2014. *HOPISICITE* est un projet porté par les laboratoires PACTE (UMR 5217) et IIAC (UMR 8177). Ce cadre est pour moi l'occasion d'aborder les questions de collections de données rétrospectives du Web. En effet, l'analyse porte sur des événements (Calais, etc.) pour la période 2013-2016. Ce projet est aussi l'occasion de prolonger la collaboration avec la BNF sur la question de l'exploitation des archives institutionnelles du Web. Il ouvre également la question de la persistance informationnelle dans le Web et les réseaux sociaux.

Ces deux projets s'inscrivent dans la continuité d'un programme de recherche centré sur les questions associées à la représentation, à des fins computationnelles et analytiques, des connaissances issues des interactions sur le Web. La dimension temporelle exprimée dans les deux contextes constitue l'un des enjeux des développements à venir de mon travail. Le temps est en effet une contrainte forte dans la réponse analytique du *juste à temps*. Il est aussi une variable clef dans la définition des *Big Data*.

L'assise disciplinaire de ce questionnement est désormais plus évidente du fait de l'évolution des pratiques de la recherche en Sciences humaines et sociales. Les subsidiarités et complémentarités disciplinaires sont mieux perçues. Le décloisonnement inscrit dans le projet du site grenoblois, contribue à des rapprochements entre laboratoires, notamment avec les laboratoires d'informatique.

C'est dans cette perspective que j'ai entrepris de me rapprocher du projet PANTEDA[278] (*Production & Analysis of Temporal Data*) porté par le LIG. Il s'agit d'un projet intégrateur pour les différentes équipes du LIG travaillant sur les *Big Data*. Le sens de ma participation est d'établir un pont entre les développements techniques de cette plateforme et les usages qui peuvent en être faits par les SHS. En effet, les collections que je réalise constituent des cas concrets d'utilisation pour ce projet. Corrélativement, les développements informatiques envisagés dans PANTEDA sont indispensables pour constituer des chaînes de traitements propres à l'analyse des collections

[277] En collaboration entre les laboratoires : (UMR 7030) LIPN Paris 13, (UMR 5194) PACTE, LIDILEM, (UMR 5224) LJK, Grenoble, (UMR) TETIS, (UMR) LIRMM (Montpellier).

[278] http://panteda.imag.fr/

massives. C'est une véritable opportunité pour mes travaux dans la mesure où les données de collections réalisées à partir de MEDIASWELL seront intégrées dans le cahier des charges de PANTEDA. C'est donc une possibilité d'aller plus loin dans la production de connaissances. En outre, ma position de traducteur me parait aller dans le sens des évolutions soulignées précédemment. Ce sera l'occasion d'engager un dialogue sur la technique qui permettra de faire exister les questionnements soulevés par une informatique appliquée aux SHS. Il ne s'agit pas seulement de considérer l'informatique comme un outil pour les SHS, mais aussi de proposer un espace réflexif sur l'évolution des pratiques scientifiques et disciplinaires dans ce champ de recherche.

Dans ce sens et sur un terme plus long, l'expérience de l'ERT UmanLab m'a familiarisé avec la logique projet de la recherche. La mise en place de structures transversales de soutien à la recherche telles que les instituts des données, dont l'Université de Grenoble Alpes s'est doté, constitue une perspective transdisciplinaire dans laquelle je souhaite m'investir. C'est, il me semble, dans ce cadre, que mon apport pour la discipline sera le plus pertinent et, je l'espère, le plus productif.

Bibliographie

A

- Abbott, A. (2009). 11. *À propos du concept de Turning Point*. In Bifurcations, édition : La Découverte, pp.187-211.
- Agamben G. (2007), *Qu'est-ce qu'un dispositif ?*, Rivages poche, petite bibliothèque, 2011, 1ere édition en italien 2007.
- Akrich, M. (1987). *Comment décrire les objets techniques*? In Techniques & Culture No.9, pp.49-64.
- Akrich, M., Callon, M., Latour, B., (1988). *À quoi tient le succès des innovations? 1: L'art de l'intéressement; 2: Le choix des porte-parole*. In Gérer et comprendre. Annales des mines (No. 11 & 12), pp.4-17.
- Akrich, M. (1989). *La construction d'un système socio-technique*. In Anthropologie et sociétés (Vol. 13, No. 2), pp.31-54.
- Angué, K. (2009). *Rôle et place de l'abduction dans la création de connaissances et dans la méthode scientifique peircienne*. Recherches qualitatives, 28(2), pp.65-94.
- Appel, V., Boulanger, H., Massou, L. (2010). *Les dispositifs d'information et de communication. Concepts, usages, objets*, Bruxelles, De Boeck.
- Appel V., Heller, T., (2010). *Dispositif et recherche en communication des organisations in les dispositifs d'information et de communication.* (Chapitre 3) Concepts, usages, objets, Bruxelles, De Boeck. pp. 39-57.
- Armatte, M., (2005). *La notion de modèle dans les Sciences Sociales: anciennes et nouvelles significations.* In Mathématiques et sciences humaines. Mathematics and social sciences, 43e année, n° 172, (4), pp.91-123.

B

- Badillo, P.-Y., Pélissier, N., (2015) *Usages et Usagers de l'information à l'ère numérique,* Revue Française des Sciences de l'Information et de la communication. N° 6. 2015. http://rfsic.revues.org/1448.
- Bastien, J.-C., Leulier, C., Scapin, D., (1998). *L'ergonomie des sites web. Créer et maintenir un service Web*, cours INRIA 28 septembre 2 octobre, ADBS, pp.111-173.
- Beaudouin, V., Fleury, S., Pasquier, M., Habert, B., Licoppe, C., (2002). *Décrire la toile pour mieux comprendre les parcours*. Réseaux, vol. 20, no 116, – FT R&D / Hermès Science Publications, pp.19-52.
- Becker, H., (2002). *Les ficelles du métier. Comment construire sa recherche en Sciences Sociales*. Paris, La Découverte.
- Bélisle, C., Bianchi J., Jourdan, R., (1999) *Pratiques médiatiques*. Paris: CNRS Editions.
- Bernard, F., Joule, R.-V., (2004). *Lien, sens et action: vers une communication engageante.* Communication et organisation, (24). https://communication organisation.revues.org/2918.
- Bernard, F., Joule, R.-V., (2005). *Le pluralisme méthodologique en sciences de l'information et de la communication à l'épreuve de la «communication engageante*. Questions de communication, (7), pp.185-208.
- Berry, D. M. (2011). *The computational turn: Thinking about the digital humanities.* Culture Machine, Vol. 12.

- Berten, A., (1999) *Dispositif, médiation, créativité : petite généalogie*, Hermès, La Revue 1999/3 n° 25, pp.31-47.
- Beuscart, J.-S., Peerbaye A. (2006) *Histoires de Dispositifs*, introduction in Terrains & travaux 2006 vol. 2 N°11, pp.3-15.
- Blandin, B., (2002) *La construction du social par les objets* Ed. PUF, Paris.
- Borges, J., Levene, M. (2000). *Data mining of user navigation patterns. In Web usage analysis and user profiling* Springer Berlin Heidelberg, pp.92-112.
- Boullier, D., (1989). *Archéologie des messageries*. Réseaux, vol. 7 N°38, pp.9-29.
- Boullier, D., (2015). *Les Sciences Sociales face aux traces du big data?* Société, opinion et répliques. *https://halshs.archives-ouvertes.fr/halshs-01141120/document.*
- Boullier, D., (2015). *Vie et mort des Sciences Sociales avec le big data*. in *Dossier : le tournant numérique… et après ?* revue Socio – la nouvelle revue des Sciences Sociales. N°4, pp.19-37.
- Bouquillon, P., Matthews, J.T., (2010). *Le web collaboratif*. Ed. PUG.
- Bourdieu, P., (1972) *Esquisse d'une théorie de la pratique*. Ed. Seuil 2000.
- Boure, R., (2002). *Les origines des sciences de l'information et de la communication: regards croisés*. Presses Univ. Septentrion.
- Boyd, D., Crawford, K., (2011). *Six Provocations for Big Data. A decade in Internet Time* - Symposium on the Dynamics of the Internet Society. Sept. 2011.
- Boyd, D., Crawford, K. (2012). *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon*. Information, communication & society, 15(5), Taylor & Francis, pp.662-679.
- Burnard, L., (2012). *Du literary and linguistic computing aux digital humanities: retour sur 40 ans de relations entre sciences humaines et informatique.* In Read/Write Book 2 – une introduction aux humanités numériques P. Mounier (dir), OpenEdition Press, pp.45-58.

C

- Cadoz, C., (1994). *Le geste canal de communication homme/machine: la communication "instrumentale"*. Technique et science informatiques, 13(1), pp.31-61.
- Caelen, J., Xuereb, A., (2007) *Interaction et pragmatique. Jeux de dialogue et de langage,* Hermès-Lavoisier, Paris.
- Callon, M., Rip, A., (1992) : *Humains, non humains : morale d'une coexistence* in *La terre outragée. Les experts sont formels !* ouvrage collectif, Sciences et société, Ed. Autrement, pp.140-156.
- Cardon, D., (2012). *Regarder les données.* Multitudes, N° 49 (2), pp.138-142.
- Cardon, D., (2015) *A quoi rêvent les algorithmes – Nos vies à l'heure des big data* Ed. La république des idées – Seuil.
- Cardon, D., Casilli, A. (2015) *Qu'est-ce que le Digital Labor?* Ina Editions.
- Casilli, A., (2010). *Petites boîtes» et individualisme en réseau.* In *Annales des Mines-Réalités industrielles* (Vol. 2010, No. 4. Eska, pp.54-59.
- Castells, M., (2007). *Communication, power and counter-power* in *the network society.* International journal of communication, 1(1), 29.
- Certeau, D., M., (1980). *L'invention du quotidien. 1. Arts de faire,* Folio Essais, Paris, Gallimard 1990.
- Chambat, P., (1994). *Usages des technologies de l'information et de la communication (TIC): évolution des problématiques.* Technologies de l'information et société, 6(3), pp.249-270.

- Chaudiron, S., Ihadjadene, M., (2010). *De la recherche de l'information aux pratiques informationnelles*. Études de communication. Langages, information, médiations, (35), pp.13-30.
- Coadic, L. Y., (1997). *Usages et usagers de l'information*. (ADBS) Nathan, Paris.
- Cooley, R., Mobasher, B., Srivastava, J., (1997). *Web mining: Information and pattern discovery on the world wide web*. In Tools with Artificial Intelligence, 1997. Proceedings, Ninth IEEE International Conference, IEEE, pp.558-567.
- Cormerais, F., Le Deuff, O., Lakel, A., Pucheu, D., (2016). *Les SIC à l'épreuve du digital et des Humanités: des origines, des concepts, des méthodes et des outils*. Revue française des sciences de l'information et de la communication, N°8, Consulté le 14/10/2016 https://rfsic.revues.org/1820.
- Courbet, D., (2011). *Objectiver l'humain?* Vol. 2 Communication et expérimentation, Hermès Science.
- Coutant, A., (2015) *Les approches sociotechniques dans la sociologie des usages en SIC*, Revue française des sciences de l'information et de la communication, N°6, consulté le 05/11/2015. http://rfsic.revues.org/1271.
- Coutaz, J., (1990). *Interfaces homme-ordinateur: conception et réalisation*. Dunod.
- Couzinet, V., (2009). *Dispositifs info-communicationnels : Questions de médiations documentaires*. (Sous la direction de). Hermes. Lavoisier, Paris, 2009.
- Cukier, K., Mayer-Schoenberger, V., (2013) *The rise of Big Data – How It's changing the way we think about the world*. The. Foreign Aff., 92, (2013) 28. http://faculty.cord.edu/andersod/The%20Rise%20of%20Big%20Data.docx.
- Cukier, K., Mayer-Schoenberger, V., (2014). *Big Data: La révolution des données est en marche*. Robert Laffont.

D

- Dahlbäck, N., Jönsson, A., Ahrenberg, L. (1993). *Wizard of Oz studies: why and how*. In Proceedings of the 1st international conference on Intelligent user interfaces, ACM, pp.193-200.
- Da Silveira, G., Borenstein, D., Fogliatto, F. S. (2001). *Mass customization: Literature review and research directions*. International journal of production economics, 72(1), pp.1-13.
- David-Ménard, M., (2008). *Agencements deleuziens, dispositifs foucaldiens* in Rue Descartes 2008/1 N° 59, pp.43-55.
- Deleuze, G., (1989). *Michel Foucault philosophe*, rencontre internationale, Paris 9-11 janvier 1988, « Des travaux », Seuil.
- Demaizière, F., (2008). *Le dispositif, un incontournable du moment* in Alsic, Vol11. N°2, http://alsic.revues.org/, pp.157-161.
- Denis, J., (2009) *Une autre sociologie des usages ? Pistes et postures pour l'étude des chaînes sociotechniques*. HAL Id: halshs-00641283 https://halshs.archives-ouvertes.fr/halshs-00641283.
- Denouël, J., Granjon, F., (2011). *Communiquer à l'ère numérique: regards croisés sur la sociologie des usages*. Presses des MINES.
- Desrosières, A., (1993) *La politique des grands nombres : histoire de la raison statistique*. Paris, La découverte.
- Ducret, A., (2011). *Le concept de «configuration» et ses implications empiriques: Elias avec et contre Weber* in SociologieS http://sociologies.revues.org/3459.

- Dumont, L., (1983). *Essais sur I'individualisme: Une perspective anthropologique sur I'idéologie modeme.* Paris: Éditions du Seuil.

E

- Elias, N., Dunning, E., (1994), *Sport et civilisation. La Violence maîtrisée* Fayard, Paris.
- Ellison, N., Boyd, D. (2007). *Social Network Sites: Definition. History, and Scholarship.* In Computer-Mediated Comunication, vol. 13 N°1, pp.16-31.
- Etzioni O. (1996). *The World-Wide Web: quagmire or gold mine?* Communications of the ACM, 39(11), pp.65-68.

F

- Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R., (2014). *Web data extraction, applications and techniques: A survey.* Knowledge-Based Systems, 70, pp.301-323.
- Flichy, P., (2004). *L'individualisme connecté entre la technique numérique et la société*. Revue Réseaux *Nouvelles réflexions sur l'internet* N°124, pp.17-51.
- Flichy, P., (2008). *Technique, usage et représentations*. Réseaux, 148(2), pp.147-174.
- Fondin, H., (2001). *La science de l'information: posture épistémologique et spécificité disciplinaire*. Vol. 38, ADBS. Documentaliste Sciences de l'Information, pp.112-122.
- Foucault, M., (1966), *Les mots et les choses.* Collection tel, Paris, Gallimard, 1966.
- Foucault, M., (1969), *L'archéologie du savoir* Collection tel, Paris, Gallimard, 1969.
- Foucault, M., (1975), *Surveiller et punir* Collection tel, Paris, Gallimard, 1975.
- Foucault, M., (1976), *Histoire de la sexualité 1 – la volonté de savoir*, Paris, Gallimard, 1976.
- Foucault, M., (1994), *Dits et écrits* Tome I : 1954-1975, sous la direction de D. Defert, F. Erwald, et la collaboration de J., Lagrange, Paris, Gallimard, 1994.
- Foucault, M., (1994), *Dits et écrits* Tome II : 1976-1988, sous la direction de D. Defert, F. Erwald, et la collaboration de J.Lagrange, Paris, Gallimard, 1994.

G

- Galinon-Mélenec, B. (2011). *L'Homme trace. Perspectives anthropologiques des traces humaines contemporaines*, Paris, CNRS éditions, série L'Homme-trace, Vol.1.
- Galinon-Mélénec, B., Zlitni, S., (2013) *Traces numériques, de la production à l'interprétation,* CNRS éditions, CNRS Alpha, 978-2-271-07239-9, pp.7-19.
- Gallezot, G., Boutin, E., Dumas, P., (2006, May). *Les Sciences de l'Information ET de la Communication: une problématique du «et»*. In XVe Congrès SFSIC, Bordeaux, Mai 2006. SFSIC.
- Garfinkel, H., (1967). *Studies in ethnomethodology*, Englewood Cliffs, N. J., Prentice-Hall.
- Gaudin, T., (1981). *Ethnotechnologie: Pour une analyse des interactions objets/sociétés.* - Cahier spécial ethnotechnologie n°2 L' EMPREINTE DE LA TECHNIQUE, Culture Technique N°4, http://documents.irevues.inist.fr/handle/2042/29789, pp.119-122.
- Gaudin, T., (2005). *La prospective.* Presses universitaires de France. Que sais-je ?
- Gavillet, I., (2010). *Michel Foucault et le dispositif : questions sur l'usage galvaudé d'un concept* (chapitre2) in *Les dispositifs d'information et de communication. Concepts, usages*, *objets*, Bruxelles, De Boeck, pp.17-38

- Gerlitz, C., Rieder, B., (2013). *Mining one percent of Twitter: Collections, baselines, sampling*. M/C Journal, N°16 vol. 2 consulté le 14/10/2016 http://journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620Rieder.
- Giddens, A., (1978). *Central Problems in Social Theory* Macmillan Londres.
- Giddens, A., (2012). *La constitution de la société: éléments de la théorie de la structuration*. Presses universitaires de France.
- Gosling, S. D., Vazire, S., Srivastava, S., John, O.P., (2004). *Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires*. American Psychologist, N°59(2), pp.93-104.
- Granjon, F., Magis, C., (2016) *Critique et humanités numériques – pour une approche matérialiste de l' "immatériel"*. In Variations (19) Revue.org. Consulté le 14/10/2016 http://variations.revues.org/748.
- Green, P., Wei-Haas, L., (1985). *The rapid development of user interfaces: Experience with the Wizard of Oz method*. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 29, No. 5). SAGE Publications., pp.470-474.
- Gysels, J., Cahen, D., (1982). *Le lustre des faucilles et les autres traces d'usage des outils en silex*. Bulletin de la Société préhistorique française, 79(7), pp.221-224

H

- Hotier, H., (2002), *Quelles perspectives pour les sciences de l'information et de la communication ?* Débat 1. In *Hommage à Robert Escarpit*, revue Communiction & Organisation N°2 (2002) (revues.org).

I

- Ibekwe-Sanjuan, F., (2014, July). *Les SIC et la technique: une filiation embarrassante*. In XIXème Congrès de la Sfsic. Penser les techniques et les technologies: Apports des Sciences de l'Information et de la Communication et perspectives de recherches, pp.1-8.

J

- Jeanneret, Y., Souchier, E., (2002). *La communication médiatisée est-elle un «usage»?* Communication et langages, 132(1), pp.5-27.
- Jeanneret, Y., (2005), *Dispositif* in *Glossaire critique de la société de l'information*. La documentation Française, pp.50-51.
- Jeanneret, Y., (2007). *Usages de l'usage, figures de la médiatisation*. Communication et langages, 151(1), pp.3-19.
- Jeanneret, Y., (2009). *La relation entre médiation et usage dans les recherches en Information-communication en France* in RECIIS, V3, N°3 Sept. 2009.
- Jenkins, H., (2006). *Convergence culture: Where old and new media collide*. New York University Press. Traduction français : *la culture de la convergence* : des médias au transmédia Armand colin (2013).
- Jones, S., (1999) *Studying the Net: Intricacies and Issues* in S. Jones (ed.), Doing Internet Research: Critical Issues and Methods for Examining the Net. Sage, pp.1-28.
- Jouët, J., (1993). *Pratiques de communication et figures de la médiation*. In Les médiations Réseaux Vol 11 N°60, pp.90-120.
- Jouët, J., (1997). *Pratiques de communication et figures de la médiation. Des médias de masse aux technologies de l'information et de la communication*. In Sociologie de la communication, Réseaux Vol 1 N°1, pp.291-312.

- Jouët, J., (2000), *Retour critique sur la sociologie des usage*s, Réseaux, 100, pp.487-521.
- Jouët, J., (2011) *Des usages de la télématique aux Internet Studies* in *Communiquer à l'ère numérique - Regards croisés sur la sociologie des usages.* Julie Denouël, Fabien Granjon, dir., Presses des Mines, Paris, pp.45-90.
- Joule, R.-V., Py, J., Bernard, F., (2004). *Qui dit quoi, à qui, en lui faisant faire quoi? Vers une communication engageante*. Psychologie sociale et communication, Dunod, pp.205-218.

K

- Katz, E., Lazarsfeld, PF., (2008). *Influence personnelle: ce que les gens font des médias.* Armand Colin.
- Kiesler, S., Sproull, L.S., (1986). *Response effects in the electronic survey.* Public Opinion Quarterly, 50, pp.402–413.
- Kosala, R., Blockeel, H., (2000). *Web mining research: A survey*. ACM Sigkdd Explorations Newsletter, 2(1), pp.1-15.

L

- Lacroix, J.-G., Moeglin, P., Tremblay, G., (1992). *Usages de la notion d'usages, Ntic et discours promotionnels au Québec et en France.* Les nouveaux espaces de l'information et de la communication, Actes du 4e congrès de la SFSIC, Lille, pp.239-248.
- Laflaquiere, J., Prié, Y., Mille, A., (2008). *Ingénierie des traces numériques d'interaction comme inscriptions de connaissances.* In 19es Journées Francophones d'Ingénierie des Connaissances (IC 2008), pp.183-195.
- Leblanc, G., (1999). *Du déplacement des modalités de contrôle,* in Hermès, La Revue 1999/3 n° 25, pp.233-242.
- Le Bodic, L., (2005) *Approche de L'évaluation Des Systèmes Interactifs Multimodaux Par Simulation Comportementale Située.* Thèse - université de Bretagne Occidentale.
- Leleu-Merviel, S., (2008) *Objectiver l'humain?* Vol. 1 *qualification, quantification,* Hermès Science.
- Le Moigne J.-L., (1977) *La Théorie du système général. Théorie de la modélisation*, version en ligne www.mcxapc.org, coll. Les Classiques du Réseau Intelligence de la Complexité, format e-book 2006 (1re éd. 1977, PUF, rééd.1986, 1990, 1994).
- Leroi-Gourhan, A., (1943) *L'homme et la matière* Sciences d'aujourd'hui - Albin Michel.
- Leroi-Gourhan, A., (1945) *Milieu et technique* Sciences d'aujourd'hui - Albin Michel.
- Licklider, J.C.R., Taylor, R.W., (1968) *The Computer as a Communication Device.* Reprinted from Science and Technology, consulté le 14/10/2016 ftp://ftp.hpl.external.hp.com/gatekeeper/pub/DEC/SRC/publications/taylor/ licklider-taylor.pdf.
- Licoppe, C., (2002) *Sociabilité et technologies de communication - deux modalités d'entretien des liens interpersonnels dans le context du déploiement des dispositifs de communication mobiles.* In Réseaux Ed. Hermès n°112-113 pp.172-210.
- Lievrouw, L.A., (2009) *New Media, Mediation, and communication study.* Information, Communication & Society. N°12 (3) http://dx.doi.org/10.1080/ 13691180802660651, pp.303-325.
- Lycett, M., (2013), *Datafication: Making Sense of (Big) Data in a Complex World,* European Journal of Information Systems, 22 (4), pp.381-386.

M

- Mallein, P., Toussaint, Y., (1994). *L'intégration sociale des technologies d'information et de communication: une sociologie des usages*. Technologies de l'information et société, 6(4), pp.315-335.
- Manovich, L., (2009). *How to follow global digital cultures, or Cultural Analytics for beginners.* Consulté le 14/10/2016 http://www.academia.edu/download/35984285/59_article_2009_how.pdf.
- Manovich, L., (2011). *Trending: The promises and the challenges of big social data.* Debates in the digital humanities, 2, pp.460-475.
- Manovich, L., (2016) *The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics*. Consulté le 21/10/2016 : http://culturalanalytics.org/2016/05/the-science-of-culture-social-computing-digital-humanities-and-cultural-analytics/.
- Marchand, P., Ratinaud, P., (2012). *L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française* Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT, pp.687-699.
- Marres, N., (2012). *The redistribution of methods: on intervention in digital social research, broadly conceived*. The sociological review, Vol60 (S1), pp.139-165.
- Mattelart, A., (2001). *Histoire de la société de l'information*. Paris, La Découverte.
- Mattelart, A., Vitalis, A. (2014). *Le profilage des populations: du livret ouvrier au cybercontrôle*. La Découverte.
- Mayer-Schönberger, V., Cukier, K., (2013). *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- McLuhan, M., (1968). *Pour comprendre les média.* Paris, Marne/Seuil.
- Merton, R.K., (1997). *Éléments de théorie et de méthode sociologique*. Paris: Armand colin.
- Merzeau, L., (2013). *Éditorialisation collaborative d'un événemen*t. Communication & Organisation, (1), pp.105-122.
- Metzger, J.-P., (2013) *L'information-documentation* in *Sciences de l'information et de la communication* sous la direction de S. Olivesi. PUG, Grenoble, pp.43-62.
- Meunier, J.-P., (1999) *Dispositif et théories de la communication : deux concepts en rapport de codétermination*. In Hermès, La Revue 1999/3 n° 25, pp.83-91.
- Meyer, T., Anis, J., (1992). *Comparer la machine à l'homme et l'homme à la machine : approche expérimentale des représentations d'une génération automatique de récit*. Langages, pp.92-105.
- Meyriat, J., (1983) : *De la science de l'information aux métiers de l'information* in *Schéma et Schématisation*, n° 19, 1983, pp.65-74.
- Miège, B., (1995). *La pensée communicationnelle*. Grenoble: Presses universitaires de Grenoble.
- Miège, B., (1996). *La société conquise par la communication. I. Logiques sociales*. Ed. PUG.
- Miège, B., (1997). *La société conquise par la communication. II. La communication entre l'industrie et l'espace public*. Ed. PUG.
- Miège, B., (2004). *L'information-communication, objet de connaissance*. De Boeck Supérieur.
- Miège, B., (2007). *La société conquise par la communication. III. Les Tic entre innovation technique et ancrage social*. Ed. PUG.
- Miège, B., Tremblay, G., (1999) *Introduction. Pour une grille de lecture du développement des techniques de l'information et de la communication*. In Sciences de la Société N°47, mai 1999.
- Mille, A., (2013) *Des traces à l'ère du Web* Intellectica, N°59-1, pp.7-28.
- Millerand, F., Proulx, S., Rueff, J., (Eds.). (2010). *Web social: mutation de la communication*. PUQ.
- Millerand, F., Proulx, S., Rueff, J., (Eds.). (2010). *Web social: mutation de la communication*. PUQ.

- MMI2 (1993), *The MMI2 Demonstrator Systems Deliverable d17 (TA2)* Esprit Project 2474 A multimodal Interface for Man Machine Interaction with Knowledge Based Systems. Ed. Wilson M.D. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.125.7058&rep=rep1&type=pdf.
- Moatti, A., (2012). *Le numérique, adjectif substantivé*. Le débat, (3), pp.133-137.
- Monnoyer-Smith, L., (2013) *Le web comme dispositif : comment appréhender le complexe?* In Barats Christine, Manuel d'analyse du web en sciences humaines et sociales, Paris, Armand Colin, pp.12-31.
- Morin, E., (1886). *La méthode 3. La connaissance de la connaissance*. Édition du seuil.
- Morin, E., (1990). *Introduction à la pensée complexe*. Éditions ESF 1990, réédition du seuil 2005.
- Mucchielli, A., (2006). *Études des communications : Le dialogue avec la technologie*. Ed. Armand Colin.

N

- Nielsen, J., (1994) *Usability Engineering,* AP Professional, Cambridge.

O

- Orlikowski, W.J., (1992). *The duality of technology: Rethinking the concept of technology in organizations*. Organization science, 3(3), pp.398-427.

P

- Paganelli, C., (2012). *Une approche info-communicationnelle des activités informationnelles en contexte de travail: Acteurs, pratiques et logiques sociales* HdR.
- Panckhurst, R., (2006). *Le discours électronique médié: bilan et perspectives. Lire, écrire, communiquer et apprendre avec Internet,* Psychologie Solal Éditeurs, pp.345-366.
- Panckhurst, R., (2007). *Discours électronique médié: quelle évolution depuis une décennie?* La langue du cyberespace: de la diversité aux normes, pp.121-136.
- Paquienséguy, F., (2006). *L'étude des usages en sic aujourd'hui : bilan et perspectives*. In Questionner les pratiques d'information et de communication. SFSIC.
- Paquienséguy, F., (2007). *Comment réfléchir à la formation des usages liés aux technologies de l'information et de la communication numériques ?* in Les Enjeux de l'information et de la communication. Vol. 1 pp.63-75.
- Paquienséguy, F., (2012). *L'usager et le consommateur à l'ère numérique* in La sociologie des usages - continuité et transformations. Sous la direction de G. Vidal. Ed. Hermès Lavoisier, pp.179-212.
- Pédauque, R., (2006), *Le Document à la lumière du numérique : forme, texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité,* C&F éditions.
- Peeters, H., Charlier, P., (1999). *Contributions à une théorie du dispositif. Le dispositif-Entre usage et concept,* in Hermès, La Revue 1999/3 n° 25, pp.15-23.
- Perriault, J., (1989). *La logique de l'usage – Essai sur les machines à communiquer*. Flammarion, Paris.
- Perriault, J., (2009). *Traces numériques personnelles, incertitude et lien social*. Hermès, La Revue, 53(1), pp.13-20.
- Perriault, J., (2015) *Retour sur la logique de l'usage*, Revue française des sciences de l'information et de la communication [En ligne], 6 | 2015, mis en ligne le 04 février 2015, consulté le 05 novembre 2015. URL : http://rfsic.revues.org/1221.

- Plantin, J.-C., Monnoyer-Smith, L., (2011) *Pour une analyse critique de l'apport heuristique et méthodologique de la recherche numérique pour les SIC*. Consulté le 14/10/2016 https://hal.archives-ouvertes.fr/hal-00641099/.
- Plantin, J.-C., Monnoyer-Smith, L. (2013). *Ouvrir la boîte à outils de la recherche numérique. Trois cas de redistribution de méthodes*. tic&société, Vol.7 N°2 (2e semestre 2013), http://ticesociete.revue.org/1527, pp.38-66.
- Polity, Y., Rostaing, H. (1997). *Cartographie d'un champ de recherche à partir du corpus des thèses de doctorat soutenues pendant 20 ans: Les sciences de l'information et de la communication en France: 1974-94*. In Actes du Colloque: Les systèmes d'informations élaborées (SFBA): France, Ile Rousse (Vol. 14).
- Poole, MS., Jackson, M., Kirsch, L., De Sanctis, G., (1998). *Alignment of system and structure in the implementation of group decision support systems*. In Conference Best Paper Proceedings, Academy of Management.
- Poole, M., Sanctis, D.G., (2004). *Structuration theory in information systems research: Methods and controversies*. The handbook of information systems research, sous la direction de M. Whitman & A. Woszczynski, Ed. Idea, pp.206-249.
- Porter, M.E., Millar, V.E., (1985). *How information gives you competitive advantage*. Harvard Business Review, july-August 1985, N° 85415, pp.149-152.
- Pourtois, J.-P., Desmet, H., (2007). *Épistémologie et instrumentation en sciences humaines*. Editions Mardaga.
- Presner, T., (2010). *Digital Humanities 2.0: a report on knowledge* consulté 14/10/2016 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.1435&rep=rep1&type=pdf .
- Proulx, S., (2001) *Usages des technologies d'information et de communication: reconsidérer le champ d'étude ?* Colloque SFSIC *:* Émergences et continuité dans les recherches en information et communication, 10-13 janvier 2001, pp.57-66.
- Proulx, S., (2005) *Penser les usages des TIC aujourd'hui : enjeux, modèles, tendances* in Lise Vieira & Nathalie Pinède, dir., Enjeux et usages des TIC : aspects sociaux et culturels (tome 1). Bordeaux : Presses universitaires de Bordeaux, 2005, pp.7-20.
- Proulx, S., (2015) *La sociologie des usages, et après ?* Revue française des sciences de l'information et de la communication [En ligne], 6 | 2015, mis en ligne le 23 janvier 2015, consulté le 05 novembre 2015. URL : http://rfsic.revues.org/1230.

Q

- Quéré, L., (1987). *L'argument sociologique de Garfinkel*. in Réseaux, volume 5, n°27, 1987. Questions de méthode, doi : 10.3406/reso.1987.1323 http://www.persee.fr/doc/reso_0751-7971_1987_num_5_27_1323, pp.97-136.

R

- Rabardel, P., Verillon, P., (1985). *Relations aux objets et développement cognitif*, in Actes des septièmes journées internationales sur l'éducation scientifique, Chamonix.
- Rabardel, P., (1995). *Les hommes et les technologies; approche cognitive des instruments contemporains*. Armand Colin. Disponible sur : https://hal.archives-ouvertes.fr/hal-01017462.
- Ratinaud, P., Marchand, P., (2012). *Application de la méthode ALCESTE à de «gros» corpus et stabilité des «mondes lexicaux»: analyse du «CableGate» avec IRaMuTeQ*. Actes des 11e Journées internationales d'Analyse statistique des Données Textuelles. JADT 2012.

- Reinert, A., (1983). *Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte*. Les cahiers de l'analyse des données, *8*(2), pp.187-198.
- Relieu, M., Salembier, P., Theureau, J., (2004). *Introduction au numéro spécial «activité et action/cognition située»*. Revue Activites. org, 1(2), pp.1-10.
- Revel, J., (2008) *Le vocabulaire de Foucault* Paris, Ellipses.
- Rieder, B., (2010). *De la communauté à l'écume: quels concepts de sociabilité pour le «web social»?* tic&société, 4(1).
- Rogers, R., (2009) *The End of the Virtual : Digital Methods* (Vol. 339). Amsterdam University Press. Consulté sur le web le 14/10/2016 https://www.researchgate.net/profile/Richard_Rogers13/publication/238579672_The_End_of_the_Virtual_Digital_Methods/links/55d4388d08ae0a34172277cd.pdf.
- Rogers, R., (2010) *Internet research question of method - A Keynote Address from the YouTube and the 2008 Election Cycle in the United States* Conference Journal of Information Technology & Politics, Vol(7) pp.241–260.
- Rogers, R., (2013) *Debanalizing Twitter: the transformation of an object of study*. In Proceedings of the 5th Annual ACM Web Science Conference pp.356-365.
- Rogers, R., (2015). *Digital methods for web research. Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*.- Wiley Online Library. Consulté le 14/10/2016 http://www.govcom.org/publications/full_list/etrds0076.pdf
- Rojot, J., (1998). *La théorie de la structuration.* Revue de gestion des ressources humaines, (26-27), pp.5-19.
- Rosa, H., (2011) *Accélération – Une critique sociale du temps*. La découverte Poche.

S

- Salaün, J.-M., (2007). *La redocumentarisation, un défi pour les sciences de l'information.* Études de communication. langages, information, médiations, (30), pp.13-23.
- Salber, D., Coutaz, J., (1993). *Applying the wizard of Oz technique to the study of multimodal systems.* In *Human-Computer Interaction* Springer Berlin Heidelberg, pp.219-230.
- Saleh, I., Hachour, H., (2012). *Le numérique comme catalyseur épistémologique* in *La théorie des industries culturelles (et informationnelles), composante des SIC* Revue Française des Sciences de l'Information et de la communication. N°1.
- Sanctis, D., G., Poole, M., (1994). *Capturing the complexity in advanced technology use: Adaptive structuration theory*. Organization science, 5(2), pp.121-147.
- Savage, M., Burrows, R., (2007). *The coming crisis of empirical sociology.* Sociology, 41(5), pp.885-899.
- Schaeffer, P., (1970). *Machines à communiquer (Vol. 2).* Éditions du Seuil.
- Searle, J.R., (1976). *A classification of illocutionary acts.* Language in society, 5(01), pp.1-23.
- Searle, J.R., (1985). *L'intentionnalité. Essai philosophique des états mentaux*, Les éditions de minuits, paris 1985.
- Serres, A., (2002). *Quelle (s) problématique (s) de la trace?*. Texte d'une communication prononcée lors du séminaire du CERCOR (actuellement CERSIC), le 13 déc. 2002,https://archivesic.ccsd.cnrs.fr/sic 00001397.
- Simondon, G., (1958). *Du mode d'existence des objets techniques* Paris, Ed. Aubier, 1989

- Srivastava, J., Cooley, R., Deshpande, M., Tan P.N., (2000). *Web usage mining: Discovery and applications of usage patterns from web data.* ACM SIGKDD Explorations Newsletter, 1(2), pp.12-23.
- Stenger, T., Coutant, A., (2010). *Les réseaux sociaux numériques: des discours de promotion à la définition d'un objet et d'une méthodologie de recherche*. In Hermes -*Journal of Language and Communication Studies*, (44), pp.209-228.
- Suchman, L., (2007). *Human-machine reconfigurations: Plans and situated actions.* Cambridge University Press.

T

- Theureau, J., (2004). *L'hypothèse de la cognition (ou action) située et la tradition d'analyse du travail de l'ergonomie de langue française.* activités, 1(2), pp.11-25.
- Thrift, N., (2005). *Knowing capitalism*. Sage.

U, V

- Van Dijck, J., (2014). *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology*. Surveillance & Society, 12(2), 197.
- Vellingiri, J., Pandian, SC., (2011). *A survey on web usage mining.* Global Journal of Computer Science and Technology, 11(4).
- Venturini, T., Cardon, D., Cointet, J.-P., (2014) *Présentation* in *Méthodes digitales Approches quali/quanti des données numériques,* Réseaux 188, no. 6: 9. doi:10.3917/res.188.0009.
- Venturini, T., Latour, B., (2010) *Le tissu social/the social fabric – traces numériques et méthodes quali-quantitatives.* Consulté août 2016.http://www.medialab.sciences-po.fr/publications/Venturini_Latour-Le_Tissu_Social.pdf.
- Vidal, G., (2012), dir. *La sociologie des usages. Continuités et transformations*. Paris : Hermès Lavoisier.

W

- Weber, M., (1971), *Économie et société*, Plon 1995.
- Weizenbaum, J. (1966). *ELIZA—a computer program for the study of natural language communication between man and machine.* Communications of the ACM, 9(1), pp.36-45.
- Wellman, B., (2004). *The three ages of internet studies: ten, five and zero years ago.* New media & society, N°6 (1), pp.123-129.
- Wieviorka, M., 2013, *L'Impératif numérique, ou La nouvelle ère des sciences humaines et sociales ?*, Paris, CNRS éditions.
- Wilson, T.D., (2000). *Human information behavior.* Informing science, vol.3, N°2, pp.49-56.
- Wirth, R., Hipp, J., (2000). *CRISP-DM: Towards a standard process model for data mining.* In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, pp.29-39.
- Wolton, D., (1997). *Penser la communication.* Champs.

X, Y, Z

- Zittoun, P., (2013) *Dispositif*, in Casillo I. avec Barbier R., Blondiaux L., Chateaureynaud F., Fourniau J-M., Lefebvre R., Neveu C. et Salles D. (dir.), *Dictionnaire critique et interdisciplinaire de la participation*, Paris, GIS Démocratie et Participation, 2013, ISSN : 2268-5863. URL : http://www.dicopart.fr/fr/ dico/dispositif.
- Zuppo, C.M., (2012). *Defining ICT in a boundaryless world: The development of a working hierarchy*. International Journal of Managing Information Technology, 4(3), pp.13-22.