Infrastructures, architectures et outils des données de la recherche

Nicolas Sauret, Jean-Luc Minel, Mélanie Bunel, Stéphane Pouyllau

Claire Scopsi, Clothilde Roullier, Martine Sin Blima-Barru et Édouard Vasseur (dir.)

Les nouveaux paradigmes de l'archive

Publications des Archives nationales

Infrastructures, architectures et outils des données de la recherche

Nicolas Sauret, Jean-Luc Minel, Mélanie Bunel et Stéphane Pouyllau

DOI : 10.4000/books.pan.7306
Éditeur : Publications des Archives nationales
Lieu d'édition : Pierrefitte-sur-Seine
Année d'édition : 2024
Date de mise en ligne : 9 février 2024
Collection : Actes
EAN électronique : 978-2-86000-390-2

Infrastructures, architectures et outils des données de la recherche

Nicolas Sauret, Jean-Luc Minel, Mélanie Bunel et Stéphane Pouyllau

1 Conceptualisés sous l'acronyme FAIR[1], les principes de la Science ouverte et de l'accès libre se sont focalisés ces dernières années sur les données de la recherche. Ce focus s'inscrit dans la continuité du mouvement d'ouverture des publications de la recherche à la fin des années 1990.

2 Comme le soulignent Cousijn, Kenall et Ganley, les scientifiques considèrent les données sur lesquelles ils étayent leurs raisonnements et leurs résultats comme des objets de recherche à part entière : « *Data citation is based on the idea that the data underlying scientific findings or assertions should be treated as first-class research objects* » (2018, p. 4).

3 Sans reprendre ici l'ensemble des principes FAIR abondamment décrits dans la littérature scientifique (Wilkinson *et al.*, 2016 ; F. W. Group, 2020 ; Wittenburg et Jong, 2020), il est utile pour la suite de ce texte de rappeler l'importance des principes de citation, de découvrabilité et d'accessibilité des données de la recherche : pouvoir citer une donnée, donner les moyens de la trouver, de la découvrir et d'y accéder. Les recommandations concernant la citation des données de la recherche dans les publications académiques ont été publiées en 2014 par la communauté FORCE11 dans la *Joint Declaration of Data Citation Principles* [JDDCP] soutenue par une centaine de sociétés savantes, d'éditeurs de revues scientifiques et d'agences de financement (D.C.S. Group, 2014). FORCE11 regroupe des chercheurs, des bibliothécaires, des archivistes, des éditeurs et des bailleurs de fonds de la recherche avec l'objectif de faciliter la transition vers une meilleure création et un meilleur partage des connaissances.

4 Si l'accès ouvert et l'ouverture des données sont parfois perçus comme des injonctions pouvant heurter les pratiques des disciplines académiques, cela fait presque deux décennies que les infrastructures de recherche françaises dédiées aux disciplines des SHS œuvrent à une traduction adaptée de ces principes dans les pratiques de ces disciplines. C'est dans ce contexte que sont nées, au tournant de 2010, les grandes plateformes de diffusion, d'archivage, d'accès et de découverte des publications et des

données numérisées, puis numériques. Combinant les pratiques anciennes de construction d'instruments de recherche par les scientifiques et la conception de dispositifs d'accès à la documentation scientifique et technique, ces plateformes ont su par ailleurs tirer profit des principes du Web et du Web sémantique (Berners-Lee et Fischetti, 1999) développés dès la fin des années 1990. C'est ainsi que le Web, cet ensemble de dispositifs sociotechniques, est devenu le socle des instruments développés par les infrastructures de recherche françaises, redéfinissant les principes de la circulation des données et des publications, puis des informations (Bermès, Isaac et Poupeau, 2013).

5 Ce chapitre[2] propose un parcours dans les infrastructures de la recherche dédiées à la conservation et à la découverte des données de la recherche en prenant en considération les pratiques de dépôt et de recherche qui se sont développées autour et avec ces infrastructures. Dans un premier temps, nous tentons de saisir ces pratiques récentes à partir d'une revue de littérature présentant quelques études de cas. Dans un second temps, nous présentons les initiatives DANS, CLARIN et ISIDORE-NAKALA, trois infrastructures représentatives de la place primordiale prise par les infrastructures de la recherche dans la gestion et la réutilisation des données de la recherche. Nous développerons plus précisément les spécificités du couple ISIDORE et NAKALA pour lequel vient d'être mené un chantier opérationnel de *fairisation* de leurs données. Enfin, en guise de conclusion, nous livrons le résultat d'un travail prospectif projetant les usages avancés que pourraient offrir des infrastructures exploitant les grands volumes de données qui y sont déposées.

Les pratiques de recherche et de dépôt des données de la recherche

6 Quelles relations entretiennent les chercheurs et les chercheuses avec les infrastructures dédiées à la gestion de leurs données ? Que ce soit pour le dépôt de leurs données ou pour la recherche de données déposées, les usagers des entrepôts ont développé des pratiques spécifiques que plusieurs études ont cherché à décrire. À la lecture de ces études, nous mettrons en exergue les principales recommandations proposées par leurs auteurs. L'examen des relations entre déposants de données, consultants des données et gestionnaires d'infrastructures de dépôt de données de la recherche nous permettra d'insister sur l'importance d'une politique de médiation, de sa définition et de sa mise en œuvre par les gestionnaires des infrastructures.

Les pratiques de dépôt pour le partage des données de la recherche

7 Les études sur les pratiques de dépôt de recherche sont relativement peu nombreuses, en raison de leur coût (temps d'analyse) et du faible taux de réponses des utilisateurs aux questionnaires ou aux demandes d'entretien. Néanmoins, quelques études de cas permettent de comprendre le comportement des chercheurs et des chercheuses vis-à-vis du partage de leurs données de recherche.

8 L'étude menée en 2018 par DANS[3] (Borgman, Darch et Golshan, 2018) présente les résultats de 9 entretiens réalisés avec des déposants. Ces échanges ont permis d'obtenir

quelques réponses sur les motivations des déposants interviewés pour partager leurs données de recherche :

- pour préserver les données dans un temps long, c'est-à-dire au-delà de la carrière professionnelle du déposant ;
- pour répondre à l'exigence de l'agence ou de l'institution qui finance le projet du déposant ;
- pour permettre à d'autres chercheurs d'exploiter leurs données.

9 L'étude met aussi en lumière les pratiques parfois inconsistantes des déposants. Par exemple, alors que la plateforme EASY gérée par DANS assigne un DOI[4] aux données déposées, les déposants ne mentionnent pas systématiquement cet identifiant dans leurs publications.

10 L'étude menée en 2017 par l'université Rennes 2 (Serres *et al.*, 2017) porte sur l'analyse des pratiques, des besoins et des attentes des chercheurs et des chercheuses des unités de recherche en sciences humaines et sociales [SHS] de l'université Rennes 2 en termes de stockage, de partage et de diffusion des données de la recherche. Cette étude approfondie, fondée sur des analyses quantitatives et qualitatives, met en exergue de nombreux points intéressants qui incitent à une réflexion globale sur les systèmes de dépôt et de partage de données, non seulement d'un point de vue technique, mais aussi épistémologique et sociologique. Contrairement à l'étude de cas précédente, les répondants n'évoquent pas comme motivation l'exigence du financeur dans le partage des données, mais mettent en avant l'indépendance des chercheurs. Par ailleurs, si l'idée de l'exploitation de leurs données par d'autres chercheurs leur paraît séduisante et altruiste sur un plan philosophique, en accord avec les principes de l'accès libre, des barrières psychologiques viennent cependant neutraliser cette motivation. En effet, l'hypothèse que les données produites dans un certain contexte et dans un but bien précis peuvent être réutilisables dans d'autres situations de recherche, voire d'autres disciplines, n'apparaît pas comme une évidence pour les chercheurs. Ce scepticisme influe énormément sur leur capacité ou leur volonté de partager leurs données. Le partage de la donnée doit alors se faire dans un contexte sécurisé, avec un contrôle sur les modalités du partage (stockage et archivage maîtrisé) afin d'assurer une réutilisation tenant compte, d'une part, de la propriété intellectuelle et, d'autre part, du contexte de production de cette donnée, en particulier l'aspect disciplinaire et le type de données concernées (qualitatives ou quantitatives). Finalement, la reconnaissance et la valorisation scientifique ne semblent pas être des facteurs de motivation dans le partage de la donnée, dans la mesure où ils sont plus efficacement couverts par la publication scientifique (articles, monographies).

11 La troisième étude de cas, réalisée en 2016 (Kim et Stanton, 2016), étudie les facteurs institutionnels et individuels qui influencent les comportements des scientifiques en matière de partage de données dans différentes disciplines scientifiques. Les auteurs s'appuient, d'une part, sur le modèle de la théorie néo-institutionnelle (Scott, 2001) et, d'autre part, sur la théorie de l'action planifiée (Ajzen, 1991). Le modèle de la théorie néo-institutionnelle identifie trois types d'influence : la contrainte régulatrice, la pression normative et la pression cognitivo-culturelle. La théorie de l'action planifiée explique le comportement d'un individu en fonction de ses intentions comportementales qui sont à leur tour influencées par son attitude à l'égard de la perception des normes subjectives. Sur la base de ces deux théories, Kim et Stanton proposent un modèle de recherche pour expliquer et prédire les comportements des scientifiques en matière de partage de données. Ils identifient deux groupes de facteurs

d'influence des comportements : les facteurs institutionnels et les facteurs individuels. Le modèle et les hypothèses développés ont été validés empiriquement en utilisant des données d'enquêtes recueillies auprès d'un panel de scientifiques appartenant à 53 disciplines (l'échantillon final comprenait 1 317 scientifiques). Les auteurs concluent leur étude par une série de recommandations :

- mettre en œuvre des politiques strictes de partage des données par les agences de financement et les revues ;
- promouvoir des normes communautaires de partage des données en s'appuyant sur les associations professionnelles ;
- développer un système d'incitation pour fournir des crédits pour le partage des données ;
- réduire les efforts nécessités par la mise en œuvre du partage des données en standardisant les protocoles de dépôts ;
- faciliter l'altruisme scientifique individuel des scientifiques en promouvant une culture altruiste de partage des données dans la communauté scientifique.

Les pratiques de recherche des données de la recherche

12 L'étude de Gregory *et al.* (2019) vise à identifier les points communs dans la façon dont les utilisateurs issus de cinq communautés de recherche (astronomie, sciences de la terre et de l'environnement, biomédecine, fouilles archéologiques, sciences sociales) recherchent et évaluent les données de recherche. Les auteurs ont collecté puis analysé la littérature sur la recherche de littérature scientifique et de données de la recherche. Cette littérature ne provenant que de la base Scopus[5], seules certaines disciplines des sciences humaines et sociales sont représentées[6]. Néanmoins, l'analyse des 400 articles de recherche collectés apporte des résultats pertinents pour notre réflexion.

13 Les auteurs notent tout d'abord que la recherche d'information et de données se fonde sur un processus identique en trois étapes : 1°) besoins utilisateurs, 2°) actions de l'utilisateur et 3°) évaluation. Ils insistent cependant sur les différences de pratiques, toutes disciplines confondues, entre recherche de publications (*Information Retrieval*) et recherche de données de la recherche.

14 Dans une seconde étude, Gregory, Cousijn et Groth (2019) articulent une analyse bibliométrique de la littérature scientifique consacrée aux pratiques de recherche avec des entretiens d'utilisateurs de la plateforme DataSearch (22 participants installés dans 12 pays) développée par Elsevier. Les auteurs présentent plusieurs résultats importants. En premier lieu, le rapport insiste sur l'importance des interactions entre un utilisateur et la communauté scientifique avec laquelle il entretient des liens. En second lieu, l'étude insiste sur l'importance d'un moteur de recherche offrant des fonctionnalités combinant différents filtres de recherche de publications et l'accès aux données de recherche associées ou citées par ces publications. En troisième lieu, le processus de recherche des données de la recherche est considéré comme un puissant levier pour mettre en œuvre des collaborations interdisciplinaires qui exploiteront au mieux les données partagées.

15 S'appuyant sur ces résultats, le rapport propose plusieurs recommandations concernant les fonctionnalités importantes que devraient offrir ces dispositifs sociotechniques :

- standardiser les métadonnées qui décrivent les données de la recherche ;
- incorporer des techniques d'enrichissement des métadonnées ;

- développer des fonctionnalités qui permettent de stimuler des collaborations autour des données ;
- développer des API[7] qui permettent d'automatiser certaines recherches dans le dépôt de données ;
- développer des outils d'interface de représentation visuelle qui permettent d'appréhender les dépôts de données selon différents points de vue (Börner et Record, 2017 ; Scharnhorst, 2015).

16 Nous ajoutons à ces recommandations l'idée de favoriser ces intersections sociotechniques en conservant et en explicitant le contexte de création et de dépôt des données, mais aussi en inférant des passerelles pertinentes entre jeux de données.

Les relations entre déposants, consultants et gestionnaires des plateformes

17 Dans leur étude, Borgman, Darch et Golshan (2018) examinent les rôles et les relations entre des déposants de données, des consultants des données et des archivistes de la plateforme DANS/EASY. Les auteurs insistent sur l'importance du rôle de médiation scientifique et technique joué par les archivistes. Ces derniers se font en effet médiateurs du libre accès aux données de plusieurs manières. L'une d'elles consiste à fournir l'infrastructure – humaine, technique et institutionnelle – facilitant le dépôt, la récupération et la gestion des données. Ils régissent les règles d'échanges entre les déposants et les consultants. Par exemple, alors que le dépôt avec des licences *Creative Commons* réduirait au minimum la médiation requise, ce modèle limiterait la capacité du DANS à acquérir des données auprès de chercheurs et chercheuses universitaires. Cette communauté a en effet exprimé qu'elle soumettrait plus volontiers des données si elle pouvait garder le contrôle sur les personnes qui y auront accès. Le verrouillage des ensembles de données oblige les consultants potentiels à s'inscrire auprès du DANS, à renseigner leur nom et à contacter directement les déposants pour demander l'accès. Le processus de demande d'accès crée un canal secondaire permettant aux déposants et aux consultants de négocier l'accès aux ensembles de données. Dans le meilleur des cas, une conversation fructueuse conduit à un partage sélectif des ensembles de données appropriés et potentiellement à une collaboration. Les données pouvant être difficiles à interpréter en dehors de leur contexte d'origine, ces relations personnelles entre déposants et consultants peuvent s'avérer essentielles à la réutilisation des données.

18 Les auteurs notent que les modèles d'utilisation des plateformes présentent les mêmes caractéristiques que les distributions de type « longue traîne » identifiées dans d'autres études sur le comportement des utilisateurs dans la recherche d'informations (Case, 2006), c'est-à-dire avec un nombre limité de grands consultants ou déposants et de nombreux utilisateurs occasionnels. Cette distribution des pratiques a plusieurs conséquences.

19 Les déposants, qui soumettent un ensemble de données une ou deux fois par an, ou peut-être une fois dans leur carrière, ont besoin d'aide au moment du dépôt pour structurer et documenter leurs données. Les documentalistes chargés des plateformes doivent vérifier les métadonnées, la documentation et l'intégrité des données pour s'assurer que les données déposées répondent aux normes minimales. Sans cette assistance par des professionnels, les données se révèlent inutilisables. Néanmoins, si les normes et la classification des métadonnées peuvent assurer un certain niveau de

découverte de base, les auteurs estiment qu'il est pratiquement impossible de normaliser les formats et les vocabulaires dans une plateforme polyvalente qui couvre plusieurs disciplines. Ils en concluent que des investissements plus importants dans les métadonnées, la documentation et les outils de recherche permettraient d'améliorer la découverte, mais des compromis sont nécessaires dans ces investissements à forte intensité de main-d'œuvre.

Infrastructures de dépôt de données de la recherche

20 Un entrepôt (ou dépôt) de données[8] est « une collection de données thématiques, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision » (Espinasse, 2021)[9]. Il existe un très grand nombre de dépôts de données de la recherche. Le catalogue Re3Data[10] référence 2 774 dépôts[11]. Il fournit un moteur de recherche et une API qui permettent de filtrer ou de parcourir le catalogue par discipline, pays, licence, etc.[12] (Buddenbohm *et al.*, 2021). De même, FAIRsharing.org[13] dénombre une liste de 1 797 dépôts[14] qu'il est possible de parcourir suivant différents critères. Chacun de ces entrepôts propose des caractéristiques spécifiques selon qu'il est géré par des institutions académiques (Harvard Dataverse, Dimonea de l'EHESS), par des institutions privées (Figshare), par des institutions disciplinaires (Pangea, CLARIN) ou multidisciplinaires (Dryad, Figshare, Mendeley, Zenodo), ou encore dédié à un seul projet (CERN Open Data Portal). Le thésaurus « Science ouverte » de l'INIST-CNRS[15] définit 7 types d'entrepôts de données ouvertes : archive ouverte, dépôt d'archive OAI[16], entrepôt agrégateur, entrepôt certifié, entrepôt disciplinaire, entrepôt institutionnel et entrepôt recommandé.

21 Il est important de souligner, à l'instar de nombreux auteurs (Borgman *et al.*, 2016 ; Karasti et Blomberg, 2017), que ces infrastructures techniques doivent être considérées comme des maillons de ce que Borgman, Darch et Golshan (2018) appellent des « infrastructures de connaissances » (*Knowledge Infrastructures)*, c'est-à-dire des « *robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds*[17] » (Edwards *et al.* 2006, p. 13), et non comme des boîtes noires dans lesquelles des données sont déposées puis recherchées.

22 Dans la suite du chapitre, nous présentons trois infrastructures de connaissances décrites selon l'angle institutionnel et selon leur attachement à une institution locale, nationale ou européenne. Les descriptions s'appuient sur leurs sites Web, sur des rapports ou des articles publiés dans des revues scientifiques et sur le rapport du COSO[18] de 2020 intitulé « Étude comparative des services nationaux de données de recherche Facteurs de réussite » (Hugo, 2020).

Une infrastructure nationale : Digital Archiving and Networked Services [DANS]

23 Fondé en 2005 par l'Académie royale néerlandaise des arts et des sciences [KNAW] et l'Organisation néerlandaise pour la recherche scientifique [NOW], DANS[19] est chargé de la gestion des données de recherche numériques en sciences humaines et sociales provenant des organisations qui l'ont précédé. DANS offre plusieurs services :

1. Electronic Archiving System (EASY)[20] pour l'archivage à long terme ;

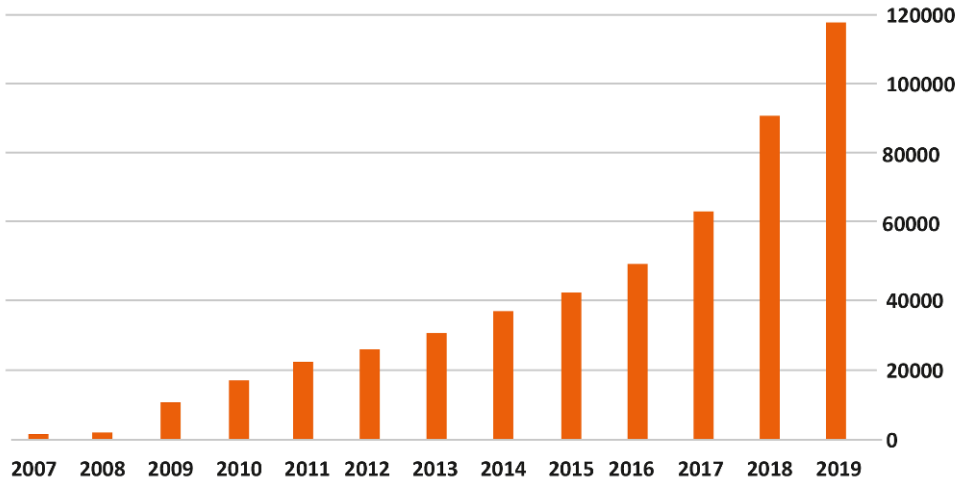
2. DataverseNL[21], un service de dépôt pour les universités, les instituts de recherche et l'enseignement supérieur ;
3. NARCIS[22], le portail national d'information sur la recherche (Doorn, 2020). DANS-EASY a reçu une certification CoreTrustSeal[23] garantissant la fiabilité des référentiels selon seize exigences. Ces exigences portent sur l'infrastructure organisationnelle, la gestion des objets numériques et la technologie sur laquelle repose DANS-EASY.

24 En 2014, DANS a créé *Research Data Netherlands*[24], une alliance visant à promouvoir les meilleures pratiques en matière de gestion et de préservation des données en partenariat avec d'autres fournisseurs néerlandais d'archives de données et d'infrastructures de recherche. DANS est par ailleurs impliqué dans de nombreux réseaux nationaux et internationaux tels que l'*European data infrastructure for scientific research* (EUDAT)[25], l'*Advanced Research Infrastructure for Archaeological Dataset Networking* (ARIADNE)[26], l'*European Open Science Cloud* (EOSC)[27] et l'*European Holocaust Research Infrastructure* (EHRI)[28].

25 DANS est organisé en trois services : « Projets et politique », « Archives et Support », « Recherche et Innovation », qui regroupent 58 personnes dont les activités sont coordonnées par un directeur. La gouvernance de DANS s'appuie sur un comité de pilotage, un comité consultatif (*advisory board*), un comité consultatif spécifique à NARCIS et un comité consultatif spécifique à DataverseNL. Le comité de pilotage de DANS supervise la gestion et le fonctionnement du réseau et des politiques menées par le directeur, ainsi que les résultats obtenus. Le comité consultatif fait part de ses recommandations en matière de stratégie et de politique générale auprès du comité de pilotage. Le comité consultatif propre à NARCIS oriente les choix de la direction de DANS au sujet du développement et du fonctionnement de NARCIS. Il est composé de représentants de sept universités et de la Bibliothèque nationale néerlandaise. Le comité consultatif spécifique à DataverseNL a pour objet de conseiller la direction de DANS sur les axes stratégiques de développement. Les 13 institutions partenaires y sont représentées.

26 En 2020, Peter K. Doorn, directeur de DANS, a publié une étude sur la montée en puissance de la plateforme EASY entre 2007 et 2019 (Doorn, 2020), dont nous reprenons ci-dessous quelques éléments.

Figure 1. Croissance du nombre de datasets dans DANS EASY, 2007-2019.



Extrait de P. K. Doorn (2020).

27 La figure 1 illustre la progression du nombre de *datasets*[29] déposés dans EASY. Après la phase de démarrage, à partir de 2012, le nombre de dépôts croît d'environ 15 à 20 % par an, puis connaît une brusque accélération avec 30 à 40 % de croissance, à partir de 2017. Doorn (2020) explique cette croissance par les conventions passées avec les universités et les institutions de recherche pour que le dépôt EASY soit utilisé comme second dépôt par ces organisations. Ces dépôts sont réalisés automatiquement sous la forme de paquets (*bulk*) échangés entre le dépôt de l'organisation et le dépôt EASY. Par ailleurs, Doorn montre que les sciences sociales représentent 30 % des dépôts et les humanités (hors archéologie) un peu moins de 10 %.

28 De ce fait, DANS répond à deux finalités :

1. la prise en charge du dépôt de données pérennes indépendamment d'une réutilisation de ces données à court ou moyen terme ;
2. la prise en charge du dépôt de données de la recherche pour répondre à des besoins de réutilisation de ces données dans l'optique de la science ouverte.

29 Doorn (2020) note une évolution importante du choix des déposants sur le type d'accès aux dépôts. Ainsi en 2012, 50 % des dépôts étaient en accès ouvert contre 70 % en 2016. L'auteur interprète cette augmentation comme un signe de la prise de conscience de l'importance d'une science ouverte.

30 Depuis 2016, DANS n'exige plus des utilisateurs un enregistrement préalable pour télécharger des données déposées en accès ouvert dans EASY. En termes d'usage, le nombre annuel d'utilisateurs enregistrés est d'environ 4 000 pour 312 472 *datasets* téléchargés entre 2007 et 2019. Ces téléchargements sont différemment répartis suivant les disciplines. Ainsi, les *datasets* en sciences sociales et dans les humanités représentent environ 4 000 téléchargements par an.

31 Dans le cadre du projet européen *Fostering Fair Data in Practices in Europe* (FAIRsFAIR)[30], le réseau DANS a collaboré avec le *Digital Curation Center* (DCC)[31] et la *Middlesex University* à la production de l'outil FAIR-Aware[32], outil d'auto-évaluation aux principes FAIR développé par DANS NL, le *Digital Curation Center* et l'université de Brême, et qui vise à sensibiliser les chercheurs et les gestionnaires de données à l'importance des principes FAIR (Hugo, 2020).

32 Les données déposées dans DANS-EASY sont accessibles *via* le moteur de recherche NARCIS. L'objectif est d'offrir aux utilisateurs les outils pour lier l'ensemble de leurs productions de manière à développer les principes d'une « *research in context* » : liens entre données, publications, chercheurs, financement et organisation.

33 DANS illustre parfaitement la notion d'infrastructure de connaissances. Cette institution interagit avec de nombreux acteurs publics et privés dans le monde entier, que ce soit des fournisseurs de moteurs de recherche, des réseaux de bibliothèques, des portails du patrimoine culturel et des sites Web qui recueillent et exploitent les données du DANS. L'infrastructure nationale se révèle ainsi être un nœud entre les infrastructures de connaissances dont elle acquiert les données et les communautés qui les consomment. Ainsi l'infrastructure déploie la technologie, définit les politiques, rédige les contrats et gère les ensembles de données qui lui sont confiés. Elle crée également des communautés en sollicitant des ensembles de données, des formations et des actions de sensibilisation. Comme ces communautés évoluent sur de longues périodes, les archives de données numériques assurent une continuité de fait articulant les différentes générations d'utilisateurs. Néanmoins, ces infrastructures sont coûteuses, nécessitent une grande quantité de temps de travail et leurs gains sont difficiles à mesurer. Leur temps de conception et de réalisation se compte en dizaines d'années et nécessite un investissement continu pour limiter une dégradation rapide (Borgman, Darch et Golshan, 2018).

Une infrastructure européenne : CLARIN

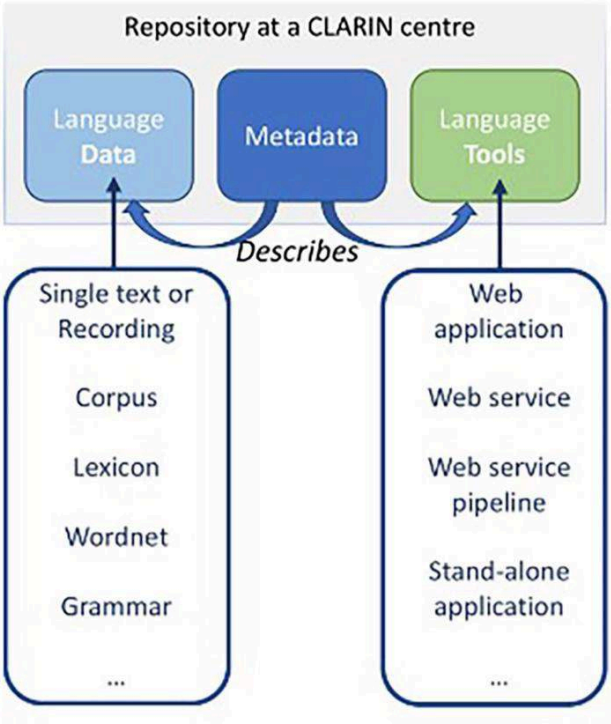
34 *Common Language Resources and Technology Infrastructure* [CLARIN][33] est un *European Research Infrastructure Consortium* [ERIC][34] créé en 2012 avec neuf membres[35] fondateurs. La tâche principale du consortium est de construire, d'exploiter, de coordonner et d'entretenir l'infrastructure de CLARIN. Il ne mène ni ne finance d'activités de recherche. CLARIN est l'une des infrastructures de recherche qui ont été sélectionnées pour la feuille de route européenne sur les infrastructures de recherche par l'ESFRI et le Forum stratégique européen sur les infrastructures de recherche. CLARIN a été créé avec le soutien financier de la Commission européenne par le biais du projet de la phase préparatoire de CLARIN (2008-2011), mais est maintenant entièrement financé par les pays participants. En 2016, CLARIN a reçu le statut de « Landmark » sur la nouvelle feuille de route. En 2017, le consortium CLARIN comprend dix-neuf pays membres et deux observateurs (dont la France) et a passé une convention avec l'université Carnegie Mellon (États-Unis). CLARIN vise une collaboration interinstitutionnelle et intersectorielle, notamment avec le secteur GLAM[36] et avec l'industrie.

35 Actuellement, CLARIN fournit un accès aux données linguistiques numériques (sous forme écrite, parlée ou multimodale) pour les chercheurs en sciences humaines et sociales. CLARIN offre également des outils avancés pour découvrir, explorer, exploiter, annoter, analyser ou combiner ces ensembles de données, où qu'ils se trouvent. Une des particularités de CLARIN est de s'appuyer sur une fédération de centres en réseau qui se distinguent selon trois types de services : des centres de dépôts de données linguistiques, des centres de services et des centres de connaissances. Les outils et les données des différents centres sont interopérables, de sorte que les collections de données peuvent être combinées et que les outils de différentes sources peuvent être articulés pour effectuer des opérations complexes (CLARIN, 2020).

36 Les centres de connaissances, au nombre de neuf en 2020[37], constituent un réseau (*Knowledge Sharing Infrastructure* [KSI]) dont une des missions est de réaliser la médiation (*the glue*) entre les infrastructures techniques et les utilisateurs. Ces centres peuvent se spécialiser dans certaines langues ou dans certaines technologies ou données. Les *workshops,* organisés par les centres et financés par le consortium CLARIN, sont considérés comme un instrument clef pour le partage des connaissances et pour le développement de nouvelles idées.

37 Les centres de dépôts peuvent se spécialiser dans une langue, une modalité (écrite ou orale), un type de données (lexicale, syntaxique, etc.) ou un type de traitement et s'engagent à être interopérables au sens où ils doivent maintenir le protocole OAI-PMH[38] pour l'échange des données. Les centres de dépôt s'engagent à respecter les principes FAIR. Le standard commun utilisé pour décrire les métadonnées est le *Component Metadata Infrastructure* [CMDI] et l'identifiant pérenne *(persistent identifier)* est un *handle* (Fig. 2).

Figure 2. Fonctionnalités générales d'un centre de dépôt CLARIN.



Extrait de Jong *et al*., 2020.

38 Les données des centres de dépôts sont moissonnées et accessibles *via* le moteur de recherche *Virtual Language Observatory* [VLO] [39] qui offre des fonctionnalités de recherche textuelles et des facettes de sélection. Néanmoins, dans leur article, les auteurs (Jong *et al*., 2020) précisent que la recherche dans un entrepôt de plus d'un million de ressources constitue un défi. Ils détaillent aussi plusieurs limitations qui tiennent aux principes mêmes de l'organisation de CLARIN. De nombreux corpus ayant été ajoutés aux dépôts nationaux ne peuvent toujours pas être identifiés dans le VLO à cause de l'absence de mots-clés ou de champs de description, ou du fait de choix idiosyncrasiques ou vernaculaires utilisés pour les dénominations. De même, des

informations sur les périodes temporelles, les annotations linguistiques et les licences d'utilisation sont absentes. L'hétérogénéité dans la granularité des *datasets* dont la taille peut varier d'un simple fichier à une archive contenant des milliers de fichiers soulève aussi des problèmes.

39 Un module de validation des métadonnées (*Curation Module*) a été développé (Ostojic, Sugimoto et Durco, 2017). Cette application contrôle un large éventail de critères (validité du schéma, présence de champs comme la langue et la disponibilité, etc.). Sur la base de ces contrôles, un score de qualité globale est calculé. Une partie très spécifique de la validation des métadonnées consiste à vérifier la validité des liens. Ces liens peuvent prendre la forme d'un identifiant de ressource de forme unique ou d'un identifiant persistant (par exemple un *handle* ou un DOI). La pratique a montré qu'environ 10 % des 5,2 millions de liens ne pouvaient être résolus par le moteur de recherche VLO. C'est pourquoi le module de curation comprend un composant spécifique qui explore régulièrement tous les liens rencontrés et stocke le résultat de l'accès à ces liens dans une base de données. Les auteurs remarquent que ce problème est générique et que plusieurs institutions développent le même type d'applications (DataCite, Europeana). Ils proposent de fédérer ces différentes bases de données dans le cadre de l'EOSC.

40 Afin de mesurer la qualité de l'infrastructure, douze indicateurs de performance *Key Performance Indicator* [KPI] ont été définis. Ces indicateurs forment un sous-ensemble des vingt indicateurs de performance définis par le groupe de travail de l'ESFRI (Report 2019).

Infrastructures FAIR : l'exemple d'ISIDORE et NAKALA

41 Le début des années 2000 a vu la production d'un très grand nombre d'études portant sur l'analyse des pratiques d'accès des chercheurs aux sources d'information et aux documents bibliographiques (publications, données de série, etc.). Dans leur étude, Ihadjadene et Chaudiron (2008) identifiaient plus d'une centaine de travaux de ce type. Dans leur conclusion, ils insistaient sur un point qui semble toujours d'actualité :

> Un moteur de recherche n'est plus simplement le « lieu » où s'apparient différentes structures cognitives dans le cadre d'interactions, mais il est considéré comme un système plus global dans lequel entrent en jeu de multiples variables : l'espace cognitif des acteurs, les caractéristiques contextuelles psychologiques, sociales et organisationnelles, ainsi que le changement des besoins d'information. Il est important d'appréhender l'usager en situation de recherche d'information de manière beaucoup plus globale que dans les modèles cognitifs et, *a fortiori*, dans l'approche système qui sous-tend encore souvent les études d'usage actuelles des moteurs de recherche (p. 29).

42 Au cours des dix dernières années, après l'arrivée de Google Scholar[40], le nombre de plateformes de recherche de documents, de données, de publications et donc d'informations a fortement augmenté (Gusenbauer 2019). L'étude menée par Know-Center (Breitfuss, Barreiros *et al.*, 2020) recense plus de 47 plateformes, incluant les plateformes privées (ResearchGate, Academia, My Science Work, etc.) ou non gouvernementales (Semantic Scholar, Dimensions, etc.). Cette étude corrobore les résultats publiés en 2016 par Lopez-Cozar, Orduna-Malea et Martin (2018), à savoir que Google Scholar est le moteur de recherche utilisé par 89 % des utilisateurs.

43 Dès la fin de la décennie 2000, la multiplication des moteurs de recherche académiques s'est doublée d'une croissance très importante de la mise à disposition de plateformes de découverte au sein des bibliothèques universitaires (Simonnot, 2012). Les *discovery tools* se sont d'ailleurs très souvent hybridés avec les moteurs de recherche académiques et les moteurs de recherche du Web (Bermès, Isaac et Poupeau, 2013 ; Gandon, Faron-Zucker et Corby, 2012). À l'échelle européenne, le développement et l'évolution de portails tels que NARCIS aux Pays-Bas, *Cultura Italia* en Italie, ou plus récemment REDIB en Espagne et en Amérique latine, ont permis aux chercheurs et chercheuses d'avoir accès de façon complémentaire et coordonnée à la littérature scientifique et aux sources de données pour les recherches dans leurs disciplines, que ce soit les publications en libre accès ou les documents sous droits, rendus accessibles *via* des API dédiées ou plus largement *via* les proxys intégrés aux portails des bibliothèques universitaires pour gérer les abonnements payants. En Europe, le développement de Driver (2006-2009) puis, depuis 2008, de la plateforme OpenAIRE[41], qui regroupe « 50 partenaires, de tous les pays de l'UE et au-delà », a offert, au cours de la décennie 2010-2019, un ensemble de « briques » pouvant être utilisées par des portails ou dispositifs de recherche de données, qu'ils soient thématiques ou disciplinaires (Manghi, Bardi et Schirrwagen, 2018).

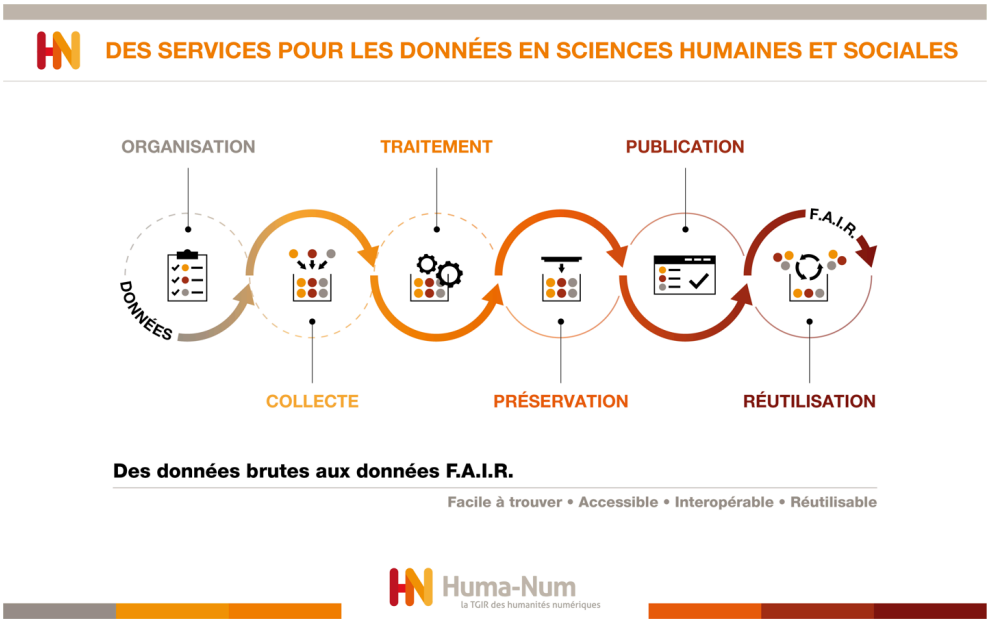
44 Dans le même temps, le développement du libre accès aux données de la recherche et aux publications (mouvement de l'*Open Access* puis de la Science ouverte) et, dans une moindre mesure, le Web sémantique (Bermès, Isaac et Poupeau, 2013), ont libéré des masses très importantes de métadonnées et de documents qui ont été intégrés à la plupart des outils de découverte sous la forme de base de données satellites. Ces dernières permettent de développer des portails associant moteurs de recherche (fondés sur l'indexation des métadonnées et du texte intégral) et outils de rebond ou d'extension par recherche fédérée vers différentes bases de données accessibles sous la forme de multiples API grâce à la proxyfication des dispositifs (Pouyllau *et al.*, 2012).

45 C'est dans ce contexte, et dès 2010, qu'ISIDORE a été imaginé et mis en œuvre (Maignien, 2011 ; Pouyllau, 2011). Au-delà de la dimension « moteur de recherche », ISIDORE s'est orienté dès le début vers l'enrichissement sémantique et la publication de métadonnées en *Linked Open Data* (Poupeau, 2016) à l'aide de référentiels scientifiques élaborés par les communautés de recherche et des bibliothèques. Et ce, dans une dimension nationale jusqu'en 2015, puis internationale avec le passage aux enrichissements multilingues à partir de 2015. L'ajout de fonctionnalités de réseau social académique dans ISIDORE est venu compléter, en 2018, un dispositif sociotechnique largement centré sur la mise en relation des savoirs avec les travaux des productions de classification et d'organisation proposées par les communautés de recherche.

46 Si NAKALA a été imaginé dans un premier temps comme un réservoir de documents et de métadonnées dans le *Linked Open Data*, sans interface Web de consultation des métadonnées et des données, ce n'est que lors de sa refonte, en 2020, que son positionnement en complémentarité de contenus avec ISIDORE a été proposé. Les deux services s'inscrivent depuis dans une interopérabilité de services (Maignien, 2011) au sein de l'écosystème d'Huma-Num. À l'image des dispositifs intégrés tels que OpenAIRE, Europeana, etc., ISIDORE et NAKALA poursuivent cette tendance en s'appuyant sur des « briques communicantes » pour faciliter leurs usages par les publics cibles. Ces briques constituent le cœur du dispositif cohérent des services mis en place par Huma-Num

pour faciliter l'accès, le signalement, la conservation et l'archivage à long terme des données de la recherche en SHS (Fig. 3).

Figure 3. Les services pour les données en SHS de Huma-Num.



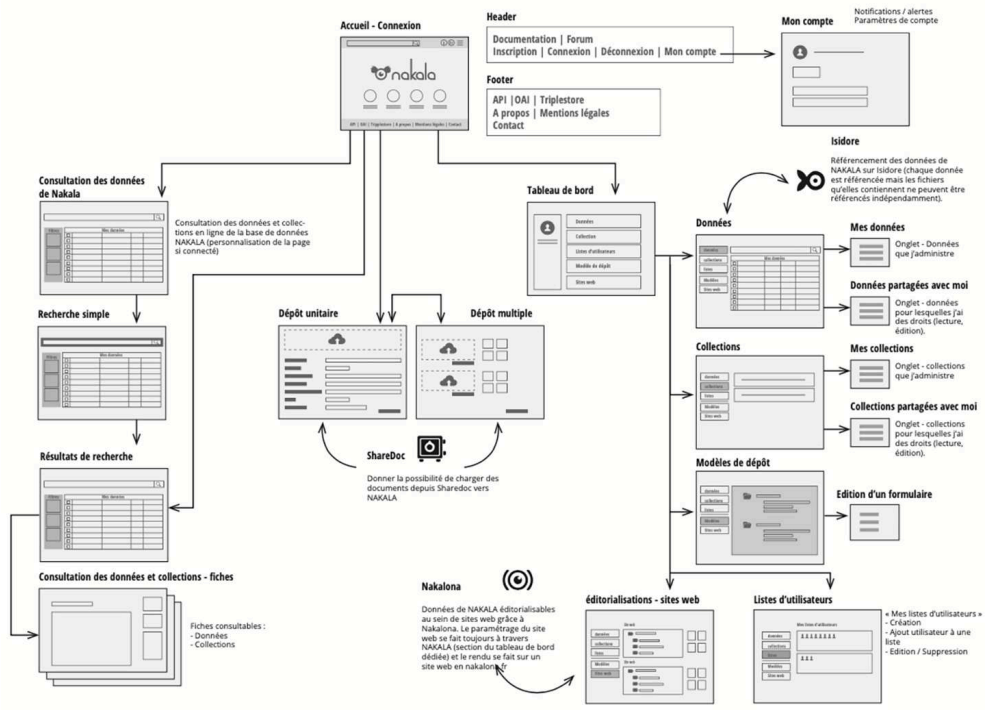
NAKALA

47 NAKALA (Fig. 4) est un service d'Huma-Num[42] permettant à des chercheurs, enseignants-chercheurs ou équipes de recherche, de partager, publier et valoriser tous types de données numériques documentées (fichiers textes, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé (Fig. 5) afin de les publier en accord avec les principes FAIR et plus largement ceux de la science ouverte (accès ouvert, immédiat et réutilisable des données publiques)[43].

Figure 4. Page d'accueil du service NAKALA.

48 L'entrepôt Nakala assure à la fois l'accessibilité aux données et aux métadonnées ainsi que leur « citabilité » dans le temps à l'aide d'identifiants stables fournis par Huma-Num et fondés sur des identifiants de type Handle (jusqu'en 2021) et DOI[44] (depuis le 19 décembre 2020). Il s'inscrit dans la politique du Web des données qui permet notamment de rendre interopérables les métadonnées, c'est-à-dire la possibilité de les connecter à d'autres entrepôts existants suivant ainsi la logique des données ouvertes et liées. Par ailleurs, il propose un dispositif d'exposition des métadonnées qui permet leur référencement par des moteurs de recherche spécialisés comme ISIDORE. La description riche, précise et harmonisée des données avec NAKALA permet d'assurer leur pérennité, de garantir leur traçabilité sur le long terme et d'encadrer leur réutilisation. L'utilisation de NAKALA a pour finalité de cibler des projets visant à publier en ligne un ensemble de données associées à des métadonnées descriptives ayant une cohérence scientifique, comme des corpus, des collections, des reportages, etc. L'objectif de NAKALA est ainsi de viser la publication de jeux de données ou d'ensemble de données ayant une valeur scientifique ou culturelle importante[45].

Figure 5. Architecture du service NAKALA.



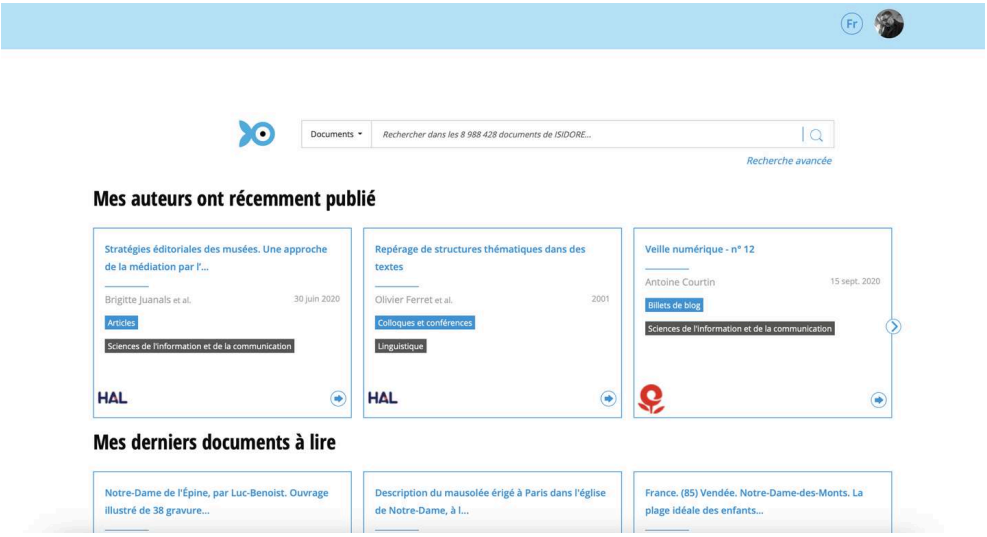
Source : L'Atelier Universel, 2021.

ISIDORE

49 Lancé le 8 décembre 2010, ISIDORE est le fruit de la collaboration du très grand équipement Adonis (Maignien, 2011) du CNRS (2007-2013), du Centre pour la communication scientifique directe [CCSD] et des sociétés Antidot, Mondéca et Sword (Pouyllau *et al.*, 2012). Il est actuellement développé et exploité par l'IR*[46] Huma-Num[47].

50 ISIDORE[48] est un moteur et assistant de recherche permettant de trouver des publications, des données numériques et profils de chercheurs et chercheuses en sciences humaines et sociales venant du monde entier (Fig. 6). Il permet de rechercher dans plusieurs millions de documents (articles, thèses et mémoires, rapports, jeux de données, pages Web, notices de bases de données, description de fonds d'archives, etc.) et des signalements d'événements (séminaires, colloques, etc.). Il propose aussi des fonctionnalités de réseau social scientifique (profil personnel, suivi d'auteurs, partage de collections bibliographiques, etc.). Il offre aussi de nombreuses fonctionnalités pour organiser sa veille scientifique (collections bibliographiques, alerte sur des requêtes, etc.).

Figure 6. Page d'accueil du service ISIDORE (anglais, espagnol et français).



51 Plus qu'un simple moteur de recherche, ISIDORE constitue une plateforme de traitement et d'enrichissement des données avec pour objectifs :

- d'offrir aux chercheurs un point d'accès unifié aux différentes ressources structurées produites dans le domaine des SHS en France ;
- d'exposer, selon les principes du *Linked Data,* les données bibliographiques structurées de la recherche en sciences humaines et sociales en France ;
- selon la logique d'une boucle de rétroaction, d'offrir les moyens aux producteurs de récupérer l'enrichissement automatique effectué par le moteur sur les données indexées.

52 ISIDORE et NAKALA constituent d'ores et déjà le « couple » central et moteur de l'écosystème Huma-Num. En 2021, Huma-Num a lancé dans le cadre du programme HNSO[49] un chantier pour renforcer ce couple et améliorer la FAIRisation des deux plateformes en travaillant sur les trois dimensions suivantes : 1°) un processus d'authentification unique, 2°) la convergence des référentiels, 3°) la procédure de moissonnage des données NAKALA par ISIDORE.

53 1. ISIDORE et NAKALA exploitent aujourd'hui le même dispositif d'authentification HumanID. Ce « hub » d'authentification offre la possibilité d'accéder en un clic à l'offre de services et d'applications de l'IR* Huma-Num en termes de stockage, de traitement, de diffusion ou encore d'exposition des données scientifiques[50]. HumanID est compatible avec l'ensemble de l'écosystème de l'enseignement supérieur et de la recherche internationale (*via* EduGAIN ou ORCID en particulier), mais aussi avec les outils les plus courants pour se connecter facilement à des services numériques. Développé avec le logiciel *open source* LemonLDAP, HumanID permet aux utilisateurs d'Huma-Num de faire des demandes d'accès aux services de l'écosystème Huma-Num et de visualiser les services connectés au Web dont ils sont ou peuvent être utilisateurs. Une évolution concrète pour l'amélioration de l'interconnexion entre ISIDORE et NAKALA consisterait à profiter de l'authentification partagée pour faciliter l'indexation des données NAKALA à partir du profil ISIDORE de l'utilisateur. Par exemple, une personne connectée et disposant de contenus « revendiqués » dans ISIDORE bénéficierait de propositions automatiques et profilées par discipline et/ou par thématique lors du dépôt de ses données dans NAKALA. Plus largement il s'agirait d'exploiter dans la base ISIDORE les informations sur les chercheurs venant des

plateformes de publications en SHS, afin de nourrir en métadonnées et de favoriser ainsi l'indexation de qualité dans NAKALA. Inversement, il serait possible d'utiliser les données déposées par un chercheur dans NAKALA pour suggérer dans ISIDORE des lectures relatives (disciplines, mots-clés, mots du titre) [Pouyllau, 2020].

54 2. ISIDORE et NAKALA exploitent les mêmes référentiels scientifico-documentaires et les mêmes auteurs. Ces référentiels communs permettent aujourd'hui de proposer à un déposant de données des labels de mots-clés (Fig. 7) et des « formes auteurs » (Fig. 8) par complétion automatique lorsque le déposant saisit ces mots-clés dans l'interface de saisie des métadonnées de NAKALA. Ce dispositif assure également une cohérence conceptuelle et une meilleure précision dans les processus de recherche d'information dans ISIDORE. Le chantier de FAIRisation propose d'améliorer le dispositif en exploitant également les URIs des labels afin d'assurer le suivi des modifications des concepts dans les référentiels et la prise en compte du multilinguisme[51].

Figure 7. Complétion automatique des labels de mots-clés présents dans les référentiels ISIDORE depuis l'interface de NAKALA.

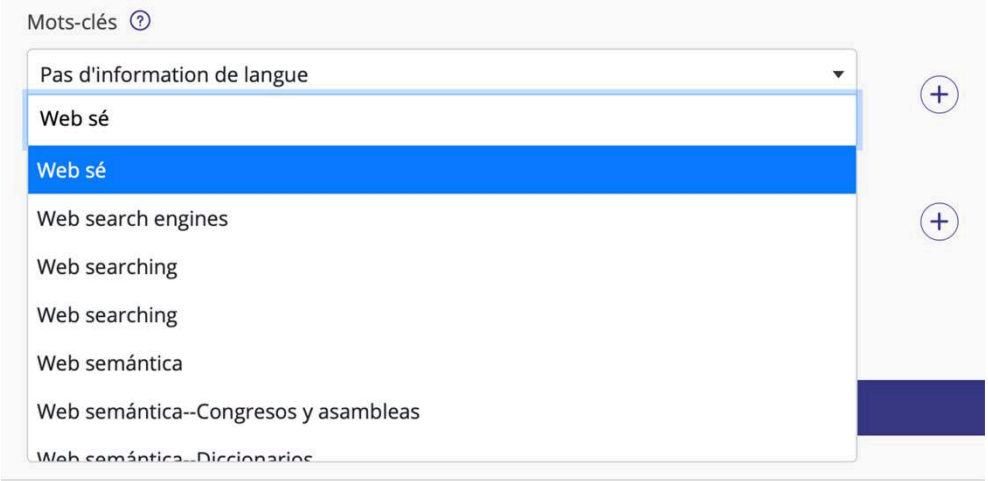
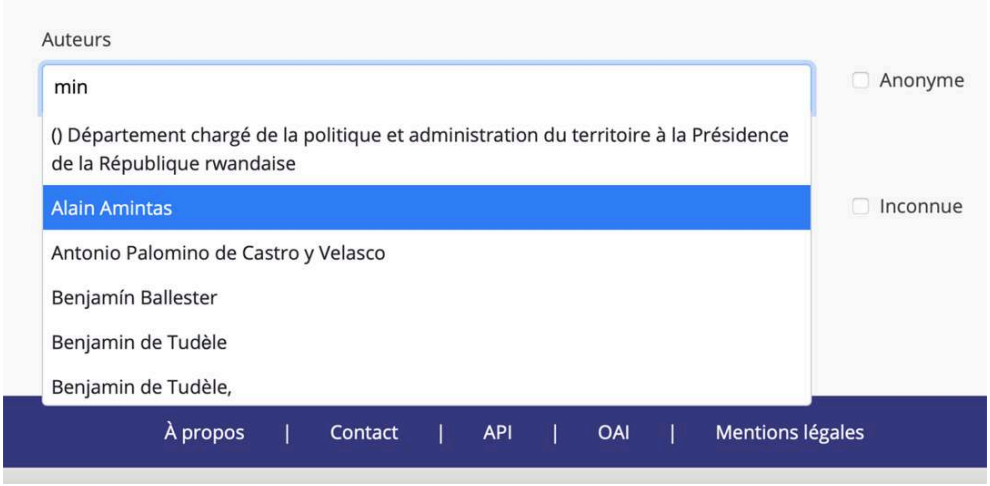


Figure 8. Complétion automatique des formes auteurs présentes dans ISIDORE depuis l'interface de NAKALA.



55 3. La création d'une collection dans NAKALA par un déposant entraîne aujourd'hui la création automatique d'un *set* dans l'exposition selon la norme OAI-PMH. Le chantier de FAIRisation prévoit qu'avec l'accord du déposant et à l'aide d'un menu accessible depuis l'interface Web de NAKALA, cette collection soit signalée à ISIDORE et automatiquement moissonnée. Alors que l'ajout de nouvelles sources ISIDORE résulte aujourd'hui d'une procédure manuelle, un ajout automatisé de nouvelles sources ISIDORE à la demande du déposant devrait très nettement améliorer la visibilité des données NAKALA.

Conclusion

56 Ce chapitre a présenté différentes infrastructures de recherche nationales et européennes dédiées à la conservation et à la découverte des données de la recherche. Nous avons mis à jour l'articulation entre infrastructures de la recherche et pratiques de la communauté académique en termes de dépôt de leurs données et de recherche au sein de ces entrepôts. Ces dernières années ont été marquées par une forte augmentation à la fois de la production de données et à la fois des usages de ces infrastructures. Usages et infrastructures sont ainsi amenés à évoluer rapidement, dans un contexte sociotechnique qui se transforme constamment comme le montre la montée en puissance des IA dans tous les domaines.

57 Sans pouvoir préjuger des orientations que prendront les différentes institutions face aux développements des pratiques, ni même celles d'Huma-Num vis-à-vis des instruments que sont ISIDORE et NAKALA, il nous semble intéressant de noter l'introduction récente des méthodes d'apprentissage machine et d'apprentissage profond qui ouvrent de nouvelles perspectives pour la gestion et le traitement des données de la recherche pour le classement, le requêtage ou la génération automatique de synthèse. Les infrastructures de recherche ont un enjeu particulier à produire des chaînes de traitement adaptées à des jeux de données à la fois très spécialisés et fortement enrichis. C'est une plus-value qui doit les différencier de la simple mise à disposition d'outils, y compris sémantiques, opération qui relève désormais davantage des DSI que des « infrastructures de recherche » au sens d'« infrastructures pour faire de la recherche ».

58 Les initiatives actuelles tentent d'exploiter le potentiel d'association des données fondée sur les principes du LOD[52], susceptible de créer des relations sémantiques entre des objets informationnels de natures différentes et provenant de plusieurs agrégateurs de données européens. Au-delà de l'Europe, le programme canadien LINCS[53] est l'un des premiers projets à mettre en œuvre une telle approche. Avec cette dernière, il devient envisageable de dépasser le portail Web de recherche documentaire pour proposer une découverte et une navigation entre concepts scientifiques, communautés de recherche et experts. L'un des enjeux de tels dispositifs sera de fournir des points d'entrées éditoriaux dans la masse des données, par exemple à partir d'une sélection de concepts disciplinaires ou en lien avec l'actualité scientifique.

BIBLIOGRAPHIE

Icek AJZEN, « The Theory of Planned Behavior », *Organizational Behavior and Human Decision Process*, 52, 2, 1991, p. 179-211.

Madeleine AKRICH, *The De-Scription of Technical Objects*, Cambridge, MIT Press, 1992.

Emmanuelle BERMÈS, Antoine ISAAC et Gauthier POUPEAU, *Le Web sémantique en bibliothèque*. Bibliothèques [Ressource électronique], Paris, Cercle de la librairie, 2013.

Tim BERNERS-LEE et Mark FISCHETTI, *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*, Harper Business, 1999.

Christine L. BORGMAN, Herbert VAN DE SOMPEL, Andrea SCHARNHORST et Henk VAN DEN BERG, « Who uses the digital data archive? An exploratory study of DANS », *Proceedings of the Association for Information Science and Technology*, 1-4, 2015, https://doi.org/10.1002/pra2.2015.145052010096.

Christine L. BORGMAN, Peter T. DARCH, Ashley E. SANDS et Milena S. GOLSHAN, « The durability and fragility of knowledge infrastructures: Lessons learned from astronomy », *Proceedings of the Association for Information Science and Technology,* 53: 1-10, 2016, http://dx.doi.org/10.1002/pra2.2016.14505301057.

Christine L. BORGMAN, Peter T. DARCH et Milena. S. GOLSHAN, « Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse », *JASIST*, 1-31, 2018, http://arxiv.org/abs/1802.02689.

Kathy BÖRNER et Elizabeth RECORD, « Macroscopes for making sense of science ». *Proceedings of the practice and experience in advanced research computing on sustainability, success and impact*, 2017, p. 64-74.

Gert BREITFUSS, Carla BARREIROS *et al.*, *Report on Stakeholder and Opportunity Analysis TRIPLE Project*, 2020, https://doi.org/10.5281/zenodo.3925662.

Stefan BUDDENBOHM, Maaike DE JONG, Jean-Luc MINEL et Yoann MORANVILLE, « Find Research Data Repositories for the Humanities – The Data Deposit Recommendation Service », *International Journal of Digital Humanities*, 2021, https://doi.org/10.1007/s42803-021-00030-7.

Daniel Owen CASE, *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior* (2nd ed.), San Diego, Academic Press, 2006.

CLARIN, *CLARIN in a nutshell*, 2020, https://www.clarin.eu/content/clarin-nutshell-0.

Helena COUSIJN, Amye KENALl et Emma GANLEy, « A data citation roadmap for scientific publishers », *Scientific Data* 5, 2018, p. 180-259.

Data Citation Synthesis Group, *Joint declaration of data citation principles, San Diego CA, FORCE11,* 2014, https://doi.org/10.25490/a97f-egyk.

Peter K. DOORN, P. K., « Archiving and Managing Research Data : data services to the domains of the humanities and social sciences and beyond : DANS in the Netherlands », *Der Archivar* 73(01), 2020, p. 44-50, https://pure.knaw.nl/portal/en/publications/archiving-and-managing-research-data-data-services-to-the-domains.

Paul N. EDWARDS, Steven J. JACKSON, Melissa K. CHALMERS *et al.*, *Knowledge infrastructures : Intellectual frameworks and research challenges*, Ann Arbor, Deep Blue, 2006, http://hdl.handle.net/2027.42/97552.

ESFRI Report Working Group, *Monitoring of Research Infrastructures Performance*, 2019, https://www.esfri.eu/sites/default/files/ESFRI_WG_Monitoring_Report.pdf.

Bernard ESPINASSE, *Introduction aux entrepôts de données,* 2021, https://pageperso.lis-lab.fr/bernard.espinasse/wp-content/uploads/2021/12/2-Intro-Entrepots-4p.pdf

European Commission, Directorate-General for Research and Innovation, *Six Recommendations for implementation of FAIR practice by the FAIR in practice task force of the European open science cloud FAIR working group*, Publications Office, 2020, https://data.europa.eu/doi/10.2777/986252.

Fabien GANDON, Catherine FARON-ZUCKER et Olivier CORBY, *Le Web sémantique : comment lier les données et les schémas sur le Web ?,* coll. « InfoPro. Management des systèmes d'information » , Paris, Dunod, 2012.

Kathleen GREGORY, « A dataset describing data discovery and reuse practices in research », *Scientific Data* 7, 2020, https://doi.org/10.1038/s41597-020-0569-5.

Kathleen GREGORY, Helena COUSIJN, Paul GROTH *et al.*, « Understanding data search as a socio-technical practice », *Journal of Information Science*, 2019, https://doi.org/10.1177/0165551519837182.

Kathleen GREGORY, Helena COUSIJN, Paul GROTH *et al.*, « Searching data: A review of observational data retrieval practices in selected disciplines », *Journal of the Association for Information Science and Technology,* 2019, https://doi.org/10.1002/asi.24165.

Michael GUSENBAUER, « Google Scholar to overshadow them all ? Comparing the sizes of 12 academic search engines and bibliographic databases », *Scientometrics,* 18, 1, 2019, p. 177-214. https://doi.org/10.1007/s11192-018-2958-5.

Catherine HUGO, *Étude comparative des services nationaux de données de recherche Facteurs de réussite, rapport du COSO*, 2020.

Madjid IHADJADENE et Stéphane CHAUDIRON, « Quelles analyses de l'usage des moteurs de recherche », *Questions de communication,* 14, 2008, p. 17-32, https://doi.org/10.4000/questionsdecommunication.604.

Franciska DE JONG, Bente MAEGAARD, Darja FIŠER, Dieter VAN UYTVANCK et Andreas WITT, « Interoperability in an Infrastructure Enabling Multidisciplinary Research: The case of CLARIN », *Proceedings LREC 2020*, p. 3406-3413, https://www.aclweb.org/anthology/2020.lrec-1.417.

Helena KARASTI et Jeanette BLOMBERG, « Studying Infrastructuring Ethnographically », *Computer Supported Cooperative Work (CSCW),* 2017, https://doi.org/10.1007/s10606-017-9296-7.

Youngseek KIM et Jeffrey M. STANTON, « Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis », *Journal of the Association for Information Science and Technology,* 67, 2016, p. 776-799, http://dx.doi.org/10.1002/asi.23424.

Emilio DELGADO LOPEZ-COZAR, Enrique ORDUNA-MALEA et Alberto MARTIN, « Google Scholar as a data source for research assessment », *Computer Science*, 2018, https://doi.org/10.31235/osf.io/pqr53.

Yannick MAIGNIEN, *ISIDORE, de l'interconnexion de données à l'intégration de services*, 2011, https://archivesic.ccsd.cnrs.fr/sic_00593320v2/document.

Paolo MANGHI, Alessia BARDI et Jochen SCHIRRWAGEN, *D7.2 – Interoperability with EOSCpilot services*, 2018, https://doi.org/10.5281/zenodo.3701434.

Davor OSTOJIC, Go SUGIMOTO et Matej DURCO, « The Curation Module and Statistical Analysis on VLO Meta-data Quality », *Selected papers from the CLARIN Annual Conference 2016*, 2017, p. 90-101.

Gautier POUPEAU, « Bilan de 15 ans de réflexion sur la gestion des données numériques », *Les petites cases (blog),* 12 octobre 2016, http://www.lespetitescases.net/bilan-reflexion-sur-la-gestion-des-donnees-numeriques.

Stéphane POUYLLAU, *ISIDORE : une plateforme de recherche de documents et d'information pour les Sciences Humaines et Sociales*, 2011, https://archivesic.ccsd.cnrs.fr/sic_00605642.

Stéphane POUYLLAU, *Classifieur de titres utilisant les données du moteur de recherche ISIDORE et l'API Keras*, 2020, https://doi.org/10.5281/zenodo.3991994.

Stéphane POUYLLAU, Jean-Luc MINEL, Shadia KILOUCHI et Laurent CAPELLI, *Bilan 2011 de la plateforme ISIDORE et perspectives 2012-2015. Comité de pilotage du TGE Adonis,* 2012, https://hal.archives-ouvertes.fr/sic_00690558.

Stéphane POUYLLAU, Mélanie BUNEL, Jean-Luc MINEL et Laurent CAPELLI, *"We": a Proposal for the TRIPLE platform*, 2020, https://doi.org/10.5281/zenodo.4032622.

Nicolas SAURET, « Design de la conversation scientifique : naissance d'un format éditorial », *Sciences du Design,* n° 8, 2, 2018, p. 57-66, https://www.cairn.info/revue-sciences-du-design-2018-2-page-57.htm.

Andrea SCHARNHORST, « Walking through a library remotely. Why we need maps for collections and how KnoweScape can help us to make them », *Les cahiers du numérique,* 11, 2015, p. 103-27.

Richard W. SCOTT, *Institutions and Organizations*, Thousand Oaks, CA, Sage Publication, 2001 [1995].

Alexandre SERRES, Marie-Laure MALINGRE, Morgane MIGNON, Cécile PIERRE et Didier COLLET, *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2 : Rapport ; Annexe 1 : Résultats de l'enquête statistique ; Annexe 2 : Croisements statistiques ; Annexe 3 : Extraits des entretiens ; Synthèse des résultats*, 2017, https://hal.archives-ouvertes.fr/hal-01635186v2.

Brigitte SIMONNOT, *L'accès à l'information en ligne. Moteurs, dispositifs et médiations*, coll. « Systèmes d'information et organisations documentaires », Hermès Lavoisier, 2012, https://archivesic.ccsd.cnrs.fr/sic_00804286.

Niels STERN, Jean-Claude GUÉDON et Thomas WIBEN JENSEN, « Crystals of Knowledge Production. An Intercontinental Conversation about Open Science and the Humanities », *Nordic Perspectives on Open Science,* 1, 2015, p. 1-24, https://doi.org/10.7557/11.3619.

Mark D. WILKINSON, Michel DUMONTIER, Ijsbrand AALBERSBERG *et al.*, « The FAIR Guiding Principles for scientific data management and stewardship », *Scientific Data* 3, 160018, 2016, https://doi.org/10.1038/sdata.2016.18.

Peter WITTENBURG et Franciska DE JONG, « State of FAIRness in ESFRI Projects », *Data Intelligence,* 2, 1-2, 2020, p. 230-237.

NOTES

1. FAIR est l'acronyme de *Findable, Accessible, Interoperable, Reusable* et se traduit en français par « Faciles à (re)trouver, Accessibles, Interopérables, Réutilisables ».

2. Ce texte reprend des parties des chapitres de l'ouvrage en édition continue « Propositions méthodologiques pour ISIDORE et NAKALA » réalisé dans le cadre du programme Huma-Num Science Ouverte [HNSO] financé par le Fonds national pour la Science ouverte.

3. Institut Digital Archiving and Networked Services (Pays-Bas).

4. *Digital Object Identifier* : identifiant pérenne et unique d'un fichier en ligne.

5. Scopus est une base de données transdisciplinaire de résumés et de citations de publications scientifiques lancée par l'éditeur scientifique Elsevier en 2004.

6. En janvier 2020, 30,4 % des titres de Scopus sont issus des sciences de la santé, 15,4 % des sciences de la vie, 28 % des sciences physiques et 26,2 % des sciences sociales. Scopus dispose d'un processus d'examen étendu et bien défini pour l'inclusion des revues ; 10 % des quelque 25 000 sources indexées dans Scopus sont publiées par Elsevier (Gregory 2020).

7. Une API (interface de programmation d'application) est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données.

8. En anglais : *Datawarehouse.*

9. D'après la définition d'Inmon (1992). Dans sa présentation, Espinasse présente les différents types de données comme suit :

- thématiques ou orientées sujet : un entrepôt de données rassemble et organise des données associées aux différentes structures fonctionnelles de l'entreprise, pertinentes pour un sujet ou thème et nécessaires aux besoins d'analyse ;
- intégrées : les données résultent de l'intégration de données provenant de différentes sources pouvant être hétérogènes ;
- historisées : les données d'un entrepôt de données représentent l'activité d'une entreprise durant une certaine période (plusieurs années) permettant d'analyser les variations d'une donnée dans le temps ;
- non volatiles : les données de l'entrepôt de données sont essentiellement utilisées en interrogation (consultation) et ne peuvent pas être modifiées (sauf certains cas de rafraîchissement).

10. *Registry of Research Data Repositories*. Registre mondial des entrepôts de données de recherche. Disponible à l'url : https://www.re3data.org.

11. Au 2 décembre 2021.

12. En appliquant les filtres « Humanities and Social Sciences », « non profit institution » et « FAIR », le catalogue recense un total de 45 infrastructures.

13. Service lié à l'université d'Oxford. Disponible à l'url : https://fairsharing.org/databases/.

14. Au 2 décembre 2021.

15. Disponible à l'url : https://www.loterre.fr/skosmos/TSO/fr/.

16. *Open Archives Initiative* (initiative pour des archives ouvertes).

17. « Des réseaux solides de personnes, de dispositifs et d'institutions qui génèrent, partagent et maintiennent des connaissances spécifiques sur les mondes humain et naturel » (Traduction des auteurs).

18. COmité pour la Science Ouverte.

19. Cette description s'appuie sur les articles de Borgman *et al.* (2015) et Doorn (2020). Disponible à l'url : https://dans.knaw.nl/nl/.

20. Disponible à l'url : https://easy.dans.knaw.nl/ui/home.

21. Disponible à l'url : https://dataverse.nl/.

22. Disponible à l'url : https://www.narcis.nl/.

23. Voir à l'url : https://www.coretrustseal.org/.

24. Disponible à l'url : https://researchdata.nl/.
25. Disponible à l'url : https://eudat.eu/european-data-initiative.
26. Disponible à l'url : https://ariadne-infrastructure.eu/.
27. Disponible à l'url : https://eosc-portal.eu/.
28. Disponible à l'url : https://www.ehri-project.eu/.
29. Un ensemble de données (*data set*) EASY est l'équivalent d'une « collection » dans la terminologie de la *Dublin Core Metadata Initiative*. Les ensembles de données sont étiquetés avec un ou plusieurs codes de classification disciplinaire.
30. Disponible à l'url : https://www.fairsfair.eu/.
31. Disponible à l'url : https://www.dcc.ac.uk/.
32. Disponible à l'url : https://fairaware.dans.knaw.nl/.
33. Cette description s'appuie sur CLARIN 2020 ; Jong *et al.*, 2020 ; Wittenburg et Jong, 2020.
34. Un ERIC est une entité juridique internationale, créée par la Commission européenne en 2009.
35. Les membres de CLARIN sont des gouvernements ou des organisations intergouvernementales.
36. Acronyme de Galleries, Libraries, Archives, Museums qui désigne le secteur des galeries, bibliothèques, archives et musées.
37. En France, le consortium CORLI est un centre de connaissance https://corli.huma-num.fr/fr/.
38. *Open Archives Initiative* Protocol *for Metadata Harvesting.* Ce protocole, développé par l'Open Archives Initiative, a pour objectif d'échanger des métadonnées entre institutions afin de multiplier les accès possibles aux documents numériques concernés.
39. Disponible à l'url : http://vlo.ChapitreInfrastructures-latest-img5.eu
40. Google Scholar a été lancé fin 2004.
41. Voir à l'url : https://www.openaire.eu/openaire-history
42. Disponible à l'url : https://www.nakala.fr.
43. Voir documentation sur NAKALA disponible en ligne : https://documentation.huma-num.fr/nakala/.
44. Disponible à l'url : https://www.doi.org/.
45. Par exemple, un fichier vidéo déposé dans NAKALA peut être inséré dans des pages Web, comme dans le cas d'un carnet de recherche *Hypothèses* (disponible à l'url : https://fr.hypotheses.org/) ou dans un webdocumentaire.
46. Appelées jusqu'en 2021 « Très Grandes Infrastructures de Recherche » [TGIR], les IR* sont les infrastructures relevant d'une politique nationale et d'un budget du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, contrairement aux IR qui sont sous la responsabilité des opérateurs de recherche.
47. Sur l'historique de l'outil, nous renvoyons le lecteur à la documentation sur ISIDORE disponible sur : https://documentation.huma-num.fr/isidore.
48. Disponible à l'url : https://isidore.science/.
49. Voir https://hnlab.huma-num.fr/blog/projets/hnso/.
50. Voir https://humanum.hypotheses.org/5754.
51. La description et l'administration des référentiels communs à ISIDORE et NAKALA sont décrits dans les chapitres Référentiels, Concepts, Définitions et Administration, Les référentiels utilisés par ISIDORE et NAKALA et Administration des référentiels utilisés dans ISIDORE de l'ouvrage HNSO « Propositions méthodologiques pour ISIDORE et NAKALA », en ligne https://hnlab.huma-num.fr/blog/2022/03/15/ouvrage-HNSO/.
52. *Linked Open Data* : « données ouvertes et liées ».
53. Linked Infrastructure for Networked Cultural Scholarship, https://lincsproject.ca

AUTEURS

NICOLAS SAURET

Maître de conférences, chercheur au Laboratoire Paragraphe, université Paris 8.

JEAN-LUC MINEL

Professeur émérite, université Paris Nanterre. Président du conseil scientifique de l'IR* Huma-Num.

MÉLANIE BUNEL

Ingénieure d'études spécialisée en ingénierie documentaire, projet Huma-Num Science Ouverte [HNSO].

STÉPHANE POUYLLAU

Ingénieur de recherche au CNRS, responsable du HN Lab.