Entrepôt de données de la recherche en SHS : un modèle de description pour l'autogestion

Eunjoo Carre, Na

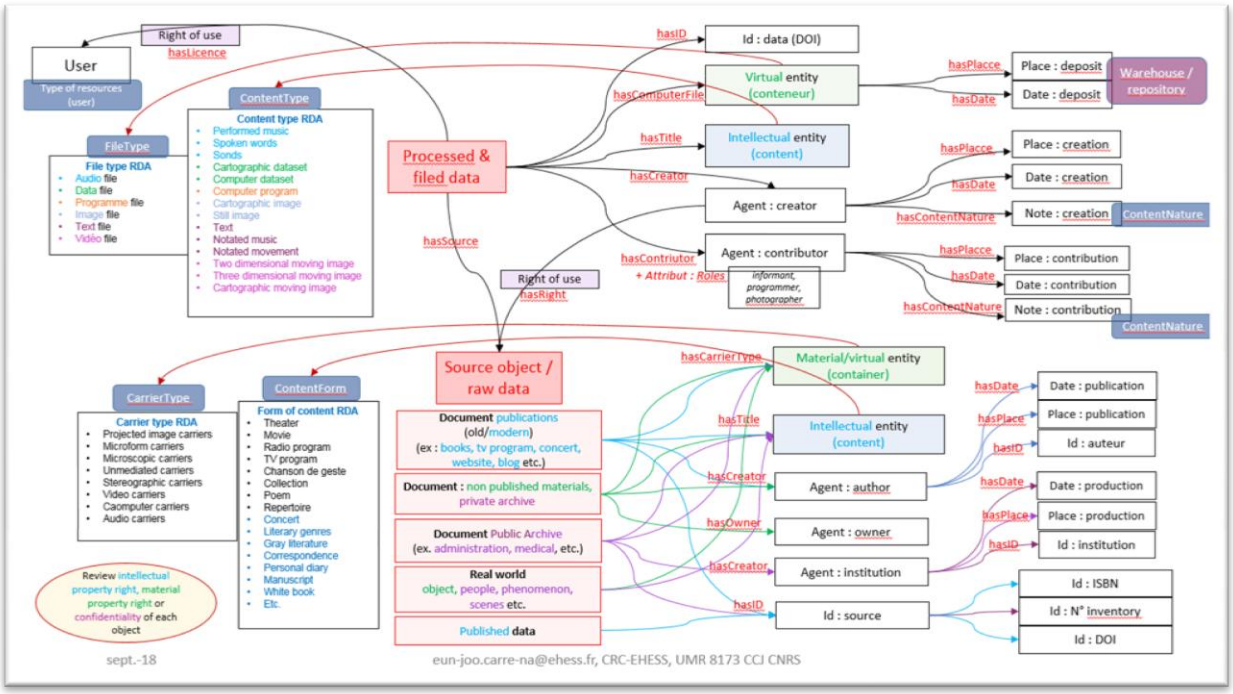
Entrepôt de données de la recherche en SHS
un modèle de description pour l'autogestion

Overseas Korean Studies Librarian Workshop
Les 15-19 octobre 2018, Bibliothèque nationale de Corée

Eun-joo Carré-Na
CRC-EHESS, UMR 8173, CCJ-CNRS

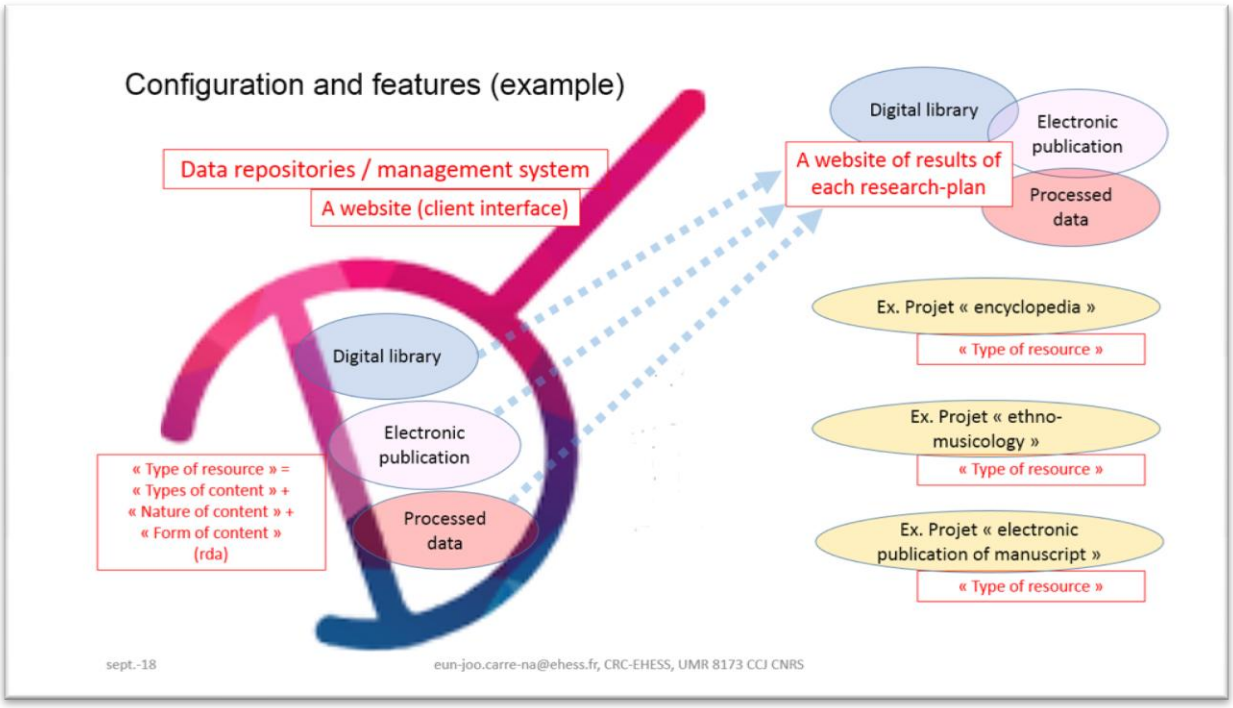
Cette communication est plutôt un appel à la réflexion et à des conseils concernant le sujet suivant : comment concevoir l'interface de dépôt des données de la recherche, pour recueillir un maximum d'informations sur les données et sur leur contexte de production, et pour que les informations recueillies prennent la forme d'un objet de gestion cohérent et exploitable, sans que le gestionnaire de document intervienne dans la procédure de dépôt (ou pour le dire autrement, sans l'étape de curation obligatoire) ?

Pour avoir des outils de communication, j'ai d'abord essayé d'encadrer et de définir des termes pour configurer les périmètres d'expression servis pour décrire le modèle que j'ai proposé au projet « Didomena » de l'EHESS.



1. Projet et missions

1.1. Projet de la conception d'un entrepôt de données de la recherche : sa position et sa politique de gestion



Dans l'ensemble du projet de soutien informatique pour la recherche à l'EHESS, ce qui est la base de mes problématiques, il faut distinguer 2 parties différentes : celle de l'entrepôt de données de la recherche, nommé « Didomena » et celle des projets de recherche, pour comprendre et respecter les caractéristiques de chaque partie : autrement dit distinguer les « généralités » et les « spécificités ».

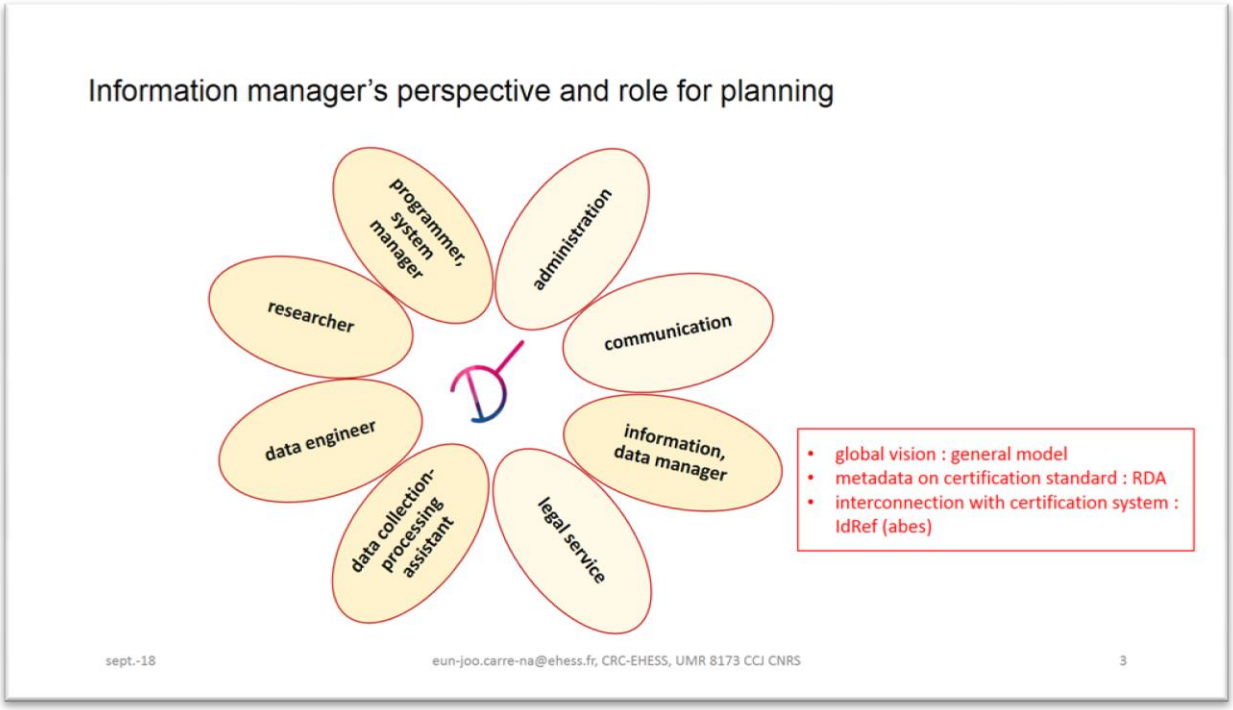
Les chercheurs, ainsi que les documentalistes (qu'ils soient gestionnaires ou non) engagés dans des projets s'intéresseraient plutôt à des problématiques propres aux projets en question, afin d'atteindre des objectifs spécifiques. En général, chaque projet de recherche concerné par Didomena est destiné à être finalisé avec la mise en place de leur propre site-web de valorisation.

Mais Didomena doit continuer à vivre et à fonctionner avec les données déposées par des projets de recherche, même après la finalisation de chaque projet. Cet entrepôt doit accueillir des données de nature très variée, issues de différents domaines de la recherche de l'EHESS.

Or Didomena a adopté une politique de « non gestion » vis-à-vis des projets de recherche, en éliminant dans l'interface de dépôt l'étape de curation. Donc, la gestion des données de chaque projet sera à la charge des responsables du projet, avec ou sans documentalistes, que ceux-ci soient gestionnaires ou pas.

En tant que documentaliste-gestionnaire, j'ai dû réfléchir pour savoir comment faire afin que Didomena puisse présenter une certaine cohérence d'identité de gestion des différentes données accueillies, sans intervenir dans la description des données au moment de leur dépôt.

1.2. Point de vue de gestionnaire pour la conception générale



Même si la gestion et l'administration du dépôt de données sont entièrement à la charge des responsables de chaque projet, je trouve nécessaire d'équiper l'entrepôt avec des moyens qui permettent de récupérer au maximum la valeur de ces données dans une cohérence normalisée. En tant que gestionnaire, j'ai essayé de garder une vision globale pour voir l'ensemble des éventuelles données, et comprendre leur variété. Puis j'ai tenté de concevoir un modèle général des « données de la recherche de l'EHESS », comme objet de gestion et comme objet de description.

De mon point de vue, les missions du gestionnaire généraliste pour un entrepôt de données sont d'abord, de chercher des **outils de descriptions**, reconnus et partagés dans le milieu professionnel de gestion documentaire, ensuite, de chercher des **moyens de connexion** nationale voire internationale, pour que cet entrepôt ne devienne pas un outil domestique isolé.

2. Outils de description : modèles et expressions

Tout d'abord, j'ai essayé de définir certains termes, pour encadrer les périmètres des expressions qui décrivent les modèles qui seront proposés, pour ne pas me perdre dans des polémiques épistémologiques interminables.

2.1. « Données de la recherche » : définition et critères

Dans le *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, publié en 2007, l'OCDE définit les « données de la recherche » comme « des **enregistrements factuels** utilisés comme sources principales pour la recherche scientifique pour **valider les résultats** de la recherche, et reconnus par **la communauté scientifique** ».

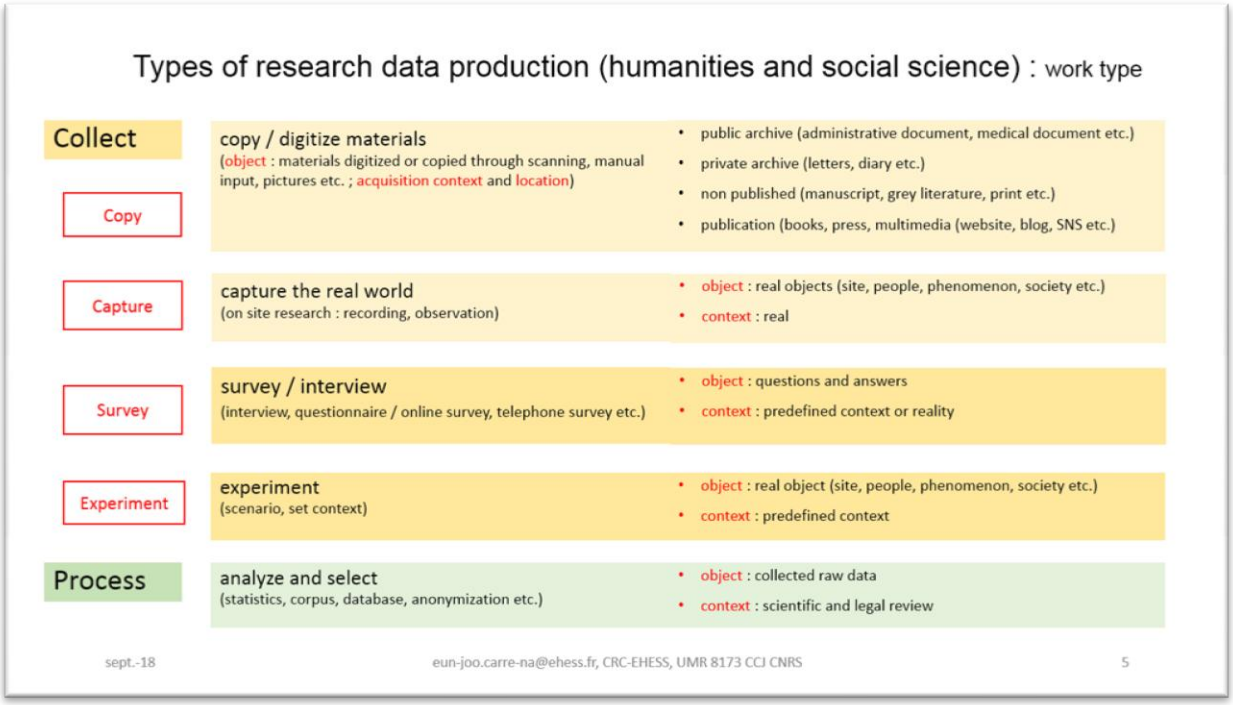
A partir de cette définition, Joachim Schöpfel et ses collègues ont tiré 4 critères : « enregistrement préalable », « nature factuelle », « lien avec la communauté scientifique » et « la finalité scientifique ».

2.2. Méthode de travail : empirique

La méthode de travail est empirique : d'abord, rassembler le plus nombreux cas de figure possibles, par tous les moyens : observation, formation et consultation ; ensuite, analyser pour déduire des éléments-atomes, composants de base auquel je reconnais un caractère commun et une utilité pour décrire et représenter des données de la recherche en science humaine et sociale (SHS).

Pour mieux comprendre ce que sont les « données de la recherche », nous (l'ensemble de l'équipe de suivi) avons d'abord recensé des exemples d'entrepôt de données, mais la plupart de ces exemples viennent de sciences de la nature, non pas de la SHS. Et puis nous avons recensé autant de projets de recherche que possible, qu'ils soient réels ou seulement potentiels, dans divers domaines de la SHS.

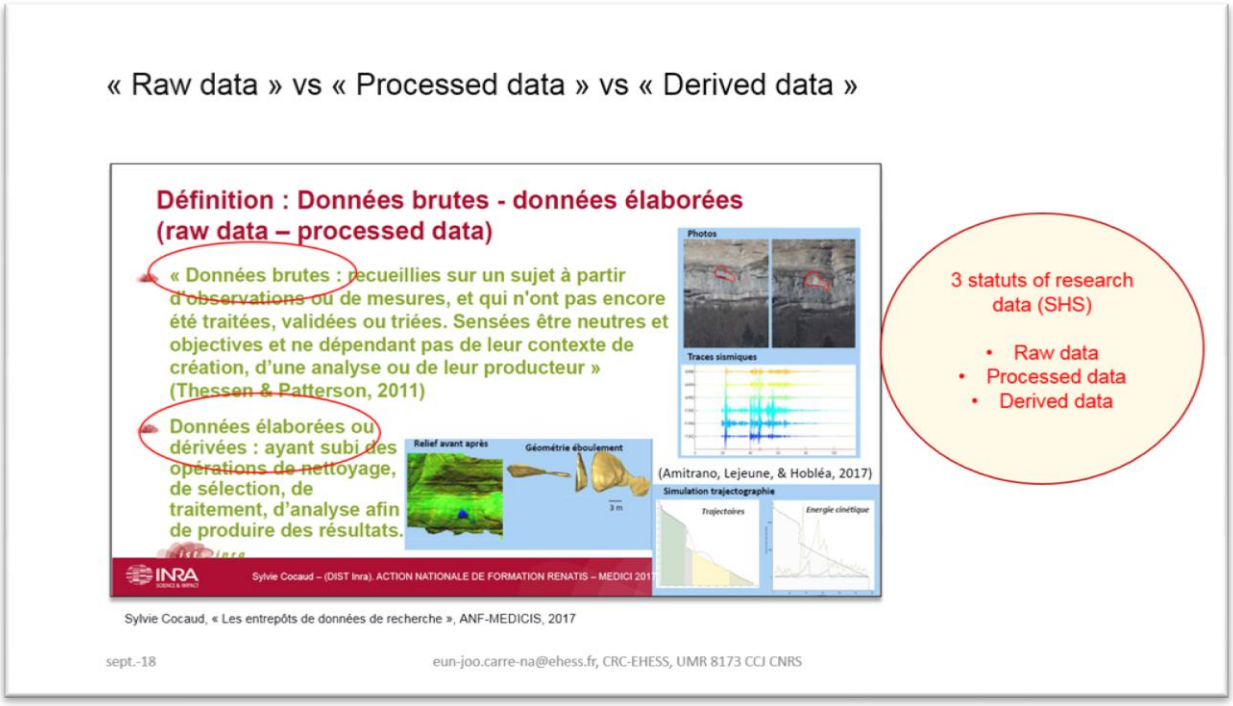
2.3. « Données de la recherche » : type de production



En analysant tous les cas de figures recensés, j'ai d'abord distingué les composants de la production : « acte », « objets », et « contexte » de production. Ensuite, j'en ai déduit ces « **types de production** » de données de la recherche, en classant diverses manières de production, sur le critère de la différence de nature de ces composants, repérée et identifiée dans chaque type.

D'abord, dans l'« **acte** de production » il y a deux étapes : « collecter » puis « traiter ». Ensuite il existe différentes formes de « **collecte** », comme « la copie, la capture, l'enquête, et l'expérimentation », et des formes de « **traitement** » comme « la sélection » et « l'analyse ». On peut remarquer que chaque type de production donne lieu à différents « objets de production » et différents « contextes de production ».

2.4. « Données de la recherche » : différents états



A la différence du document dit « classique », dont la forme finale est l'objet habituel des gestionnaires de documents, les données de la recherche ont une vie, elles évoluent avec l'avancement du projet scientifique : d'où les différents états d'une donnée de la recherche. Nous avons déjà vu les 2 actes de production « collecter » et « traiter ». Ces étapes créent d'abord 2 différents états de donnée : la « donnée brute » et la « donnée élaborée ».

Je me suis basée sur la présentation de Sylvie Cocaud, la plus récente que j'ai pu trouver. Elle a signalé et défini ces 2 différents états de donnée. Dans sa définition, elle a traité comme des états identique ou similaire, la « donnée élaborée » et la « donnée dérivée ». Mais son travail se situe dans la cadre des sciences de la nature, non pas en SHS. Alors, je les ai interprétées un peu différemment, en les adaptant à la situation de SHS, surtout en remarquant leur différente position dans le cycle de vie.

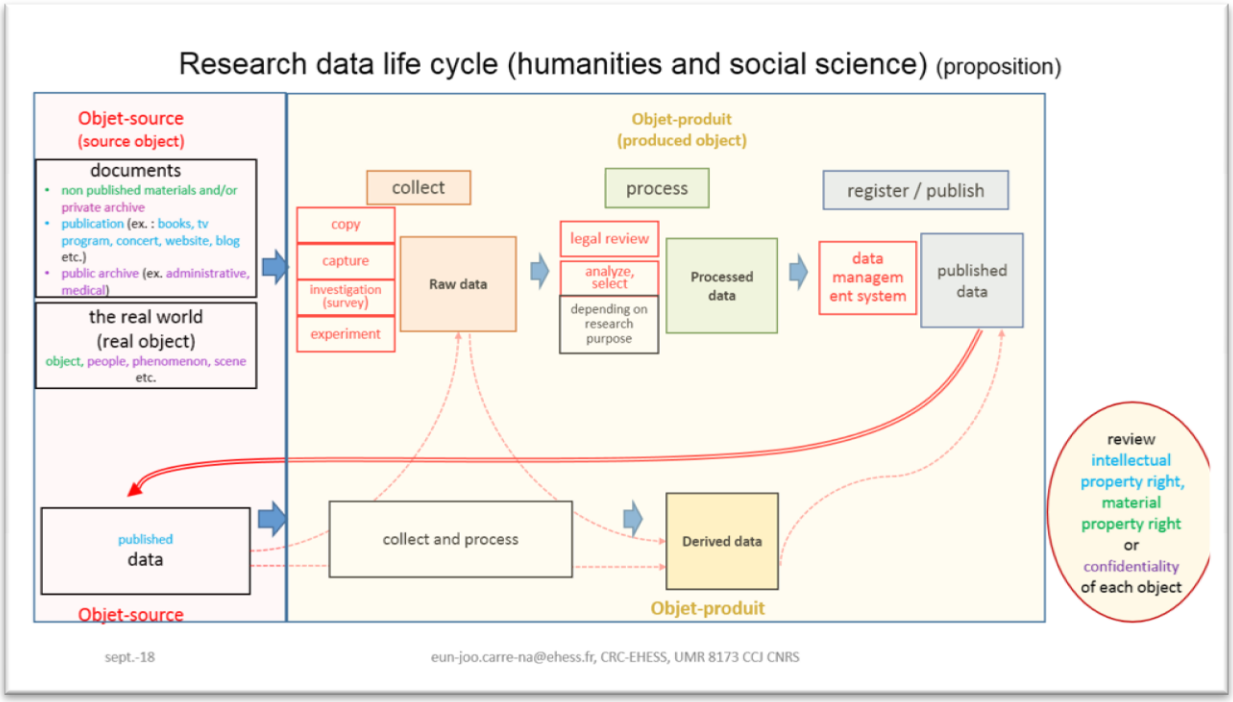
Les « **données brutes** » sont des données collectées par des chercheurs, mais pas encore avec des idées précisées sous forme d'un projet scientifique. En général, les chercheurs conservent ces données brutes en vrac dans un espace privé.

Quand ils conçoivent un projet de recherche, ils les trient et sélectionnent selon les objectifs de leur projet. Il s'agit du début de « **données élaborées** » qui continuent à être élaborées au fur et à mesure de l'avancement du ce projet, sans violer leur nature factuelle.

Sylvie Cocaud a défini en même temps les « données élaborées » et les « **données dérivées** ». Ces 2 dernières figurent dans le processus de production en tant qu'un « produit fini ». Mais je souhaiterais distinguer ces deux états de donné, en soulignant la différence de leur objet-source, et en repérant différente position dans le cycle de vie. Je définirais les « données dérivées » comme des données produites à partir des données de la recherche déjà élaborées et déjà publiées dans le cadre d'un autre projet de recherche.

Pour récapituler dans le cycle de vie de la donnée de la recherche, il y a selon moi 3 états de données : « brute », « élaboré » et « dérivé », au lieu de 2, qui peuvent être visualisés comme le tableau suivant.

2.5. « Données de la recherche » : cycle de vie



Le cycle de vie est important pour les données de la recherche, qu'il s'agisse de sciences de la nature ou de sciences humaines et sociales, parce qu'à chaque étape de vie l'état de la donnée peut se modifier.

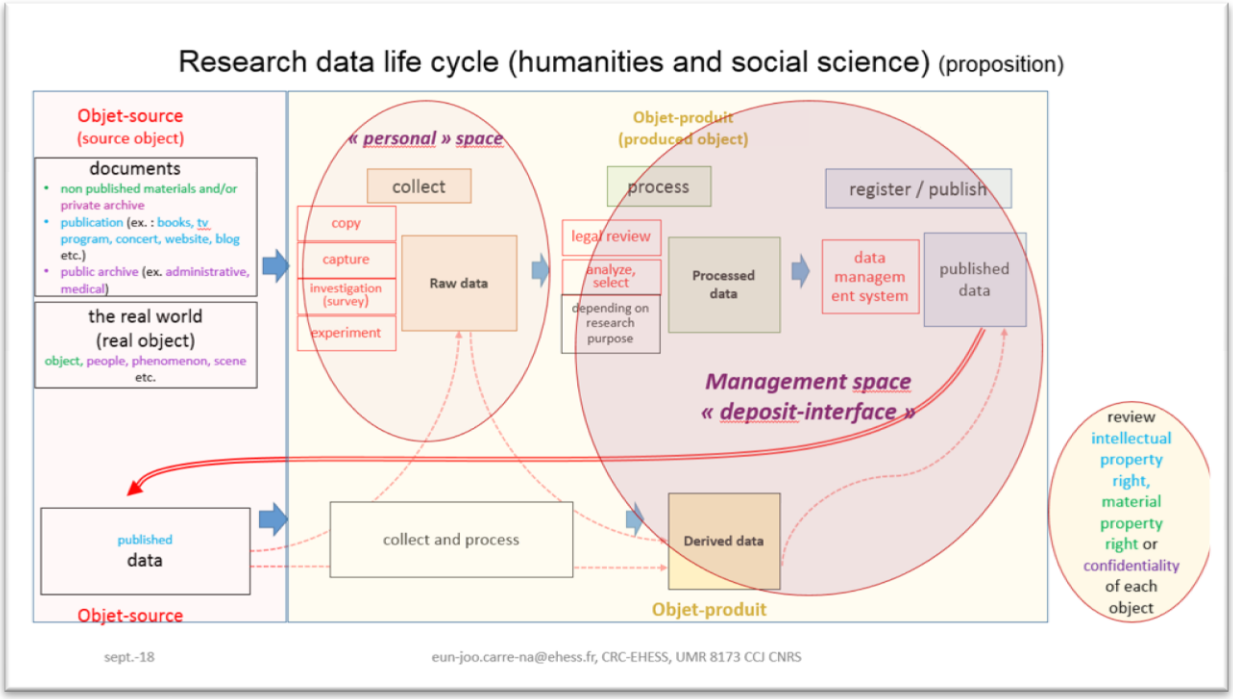
Les gestionnaires sont habitués aux documents dans un état de production fini et stable, présentés sous une forme plus ou moins conventionnelle. Donc, leur contexte de production n'était pas forcément un élément important à gérer, ni à décrire, jusqu'à présent.

Dans le cycle de vie de données de la recherche, il faut considérer aux 2 extrémité du processus de production, les 2 objets de production : l'« objet-source » et l' « objet-produit ».

En SHS, l'objet-source des données de la recherche sont souvent des « documents et archives » selon l'expression habituelle et conventionnelle (ex. domaine « histoire »), ou bien des éléments réels du monde que souvent les anthropologues, ethnologues et sociologues observent et capturent comme références.

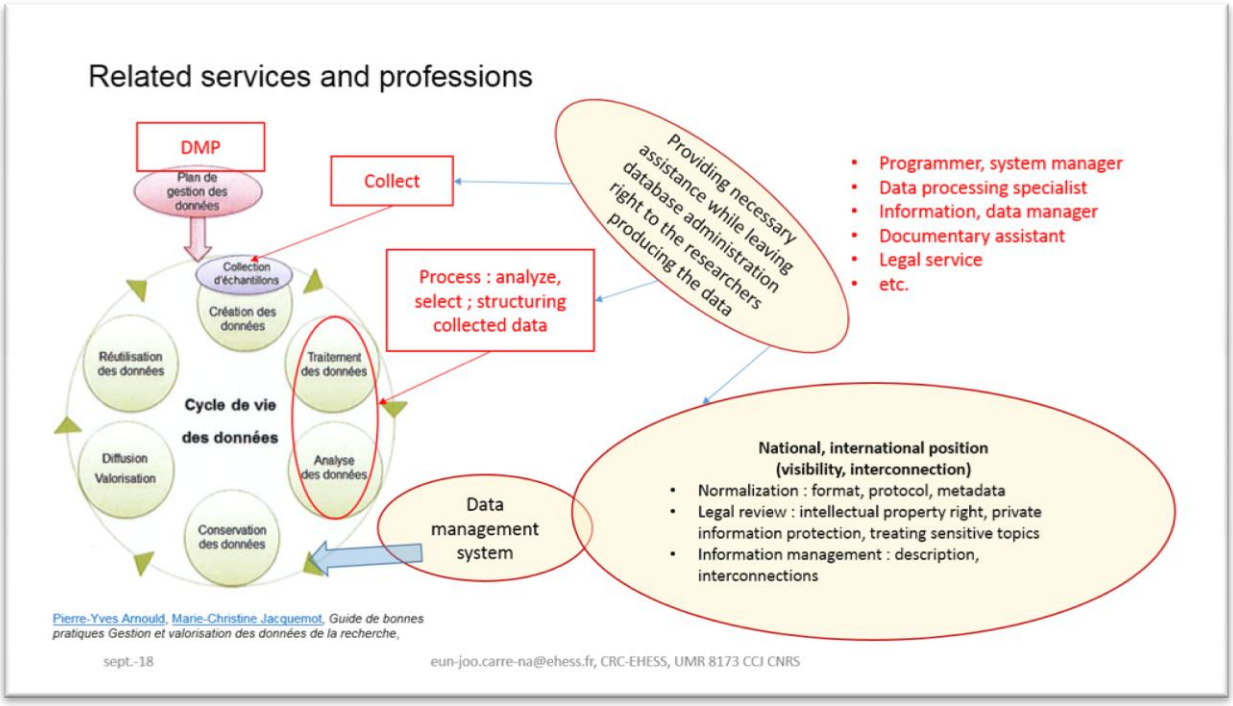
Je pense qu'une bonne description de l'objet-source de chaque donnée déposée, qui tiendrait bien compte de ces différents états, permettrait de générer automatiquement le cycle de vie des données en question, ainsi que sa citation.

2.6. Objet de la gestion de l'entrepôt de données



L'objet de la gestion par l'entrepôt des données de la recherche sont des données en tant qu'objets produits par et pour un projet scientifique, c'est-à-dire des données élaborées ou dérivées prêtes à déposer. Les données en état brut, stockées dans un espace privé ne sont pas l'objet de gestion de l'entrepôt. Le changement d'état entre « brut » et « élaboré » ne signifie pas forcément la modification du contenu, et encore moins la modification de sa nature factuelle, mais simplement un changement de statut, selon l'attachement à un objectif scientifique, avec ou sans transformation de la forme du contenu.

2.7. Services et métiers : les missions du documentaliste- gestionnaire



Dans le guide de l'INRIA de 2016, le cycle de vie est exprimé différemment, mais le résultat revient au même. Dans ce guide pratique pour les sciences de la nature, les auteurs ont séparé le « traitement » (ex. nettoyage des extraits d'échantillon) et l'« analyse ». Mais en SHS où on ne peut pas distinguer clairement l'« analyse » du « traitement » par sélection, j'ai assemblé ces deux actes en un seul : « traiter » comme « analyse pour la sélection ». Puis j'ai vu la position de l'entrepôt entre le « traitement » et la « conservation puis diffusion ».

Chaque corps de métier pourrait définir ses propres missions vis-à-vis de l'entrepôt, pour apporter un service de soutien adapté à chaque étape de vie des données.

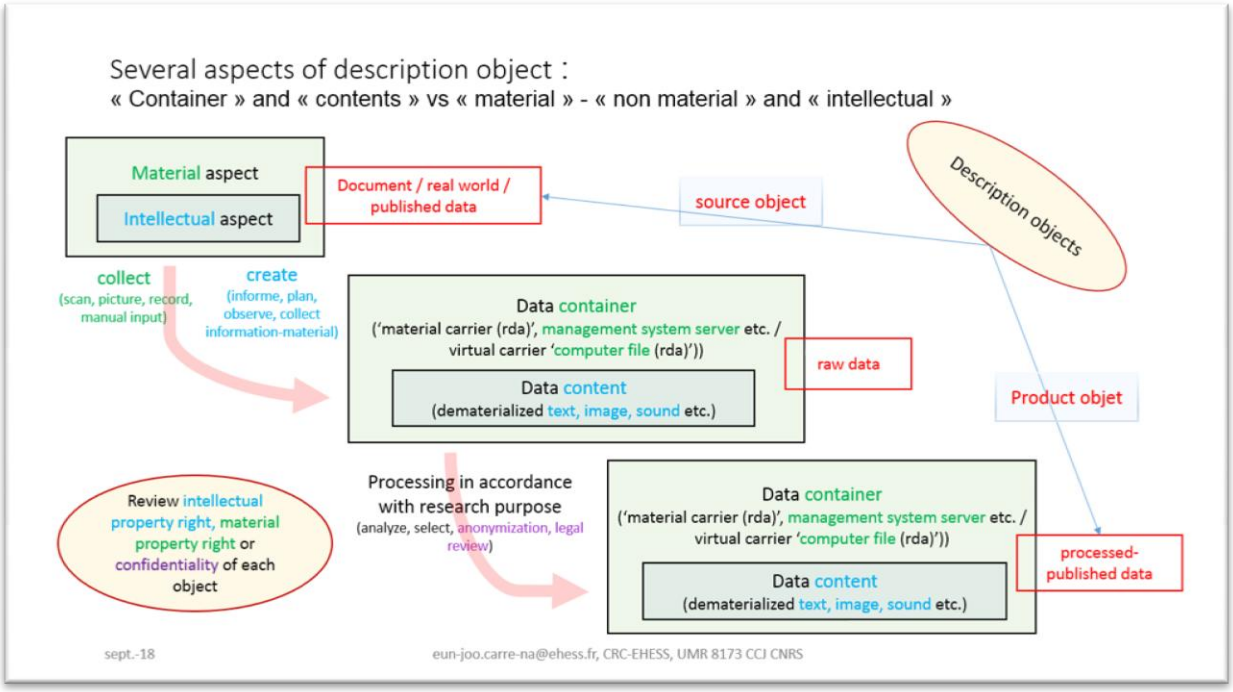
Les missions de l'entrepôt de données de la recherche, telles que je les comprends, pour les documentalistes-gestionnaires, c'est d'aider les concepteurs de données pour que leurs données déposées puissent bien **s'intégrer** dans le milieu scientifique au niveau national et international, en leur offrant une **visibilité** et des points d'**interconnexion**. Cette insertion sera réalisée par la « normalisation », la « révision juridique » et l'« enrichissement documentaire », en satisfaisant le protocole FAIR (Findable, Accessible, Interoperable, Reusable).

On peut les aider pour la « **normalisation** », en guidant les déposants vers l'utilisation des **formats standards.** On peut imaginer une fonctionnalité de **conversion en format standard**. Egalement, la « normalisation » de données peut être guidée à travers les champs de description conçus avec des **métadonnées** internationalement reconnues. De même, par le champ d' « outil d'exécution », on peut guider et/ou offrir le logiciel nécessaire à la consultation des données déposées-publiées.

On peut également apporter un soutien pour la « **révision juridique** » : en leur faisant vérifier la réglementation en terme de propriété intellectuelle, sur les sujets sensibles, la protection des informations personnelles, etc., avec des champs obligatoires de « droit » et de « licence » au moment de dépôt.

Pour l'« **enrichissement documentaire** », on devrait guider les déposants pour qu'ils puissent donner un maximum d'informations non seulement sur les données en question, mais aussi sur leur contexte de production, parce que pour les données de recherche, leur contexte de production est une information cruciale et indispensable pour prouver leur « nature factuelle ». De plus en faisant indiquer les identifiants de chaque objet-source, on garde la trace documentaire, et on crée une potentialité pour des interconnexions interne et externe, même si on ne laisse pas toujours une trace visible par le public. En retour, tous ces éléments d'interconnexion donneront aux données déposées une meilleure visibilité.

2.8. Modèle de description : différents aspects des objets



Pour décrire des données de la recherche comme ressource du web sémantique, il vaudrait mieux être conscient des différents aspects des objets de production. Pour ces objets de description, les gestionnaires devraient distinguer les aspects « matériel » et « intellectuel », désignés par convention comme « conteneur » et « contenu » des données. Le « conteneur » est de l'aspect matériel, par rapport au « contenu », l'aspect intellectuel et abstrait.

On voit que ce résultat d'analyse rejoint plus ou moins le modèle FRBR à la base de RDA. Ces deux termes renvoient respectivement aux expressions de RDA : « support matériel » et « contenu » ou « œuvre/ expression ».

3. Choix des outils : RDA et IdRef

D'après ma définition du point de vue des gestionnaires généralistes de Didomena, évoqué plus haut, j'ai essayé de donner une cohérence structurelle aux données déposées, en cherchant des outils de description reconnus, et des moyens de connexion. Et j'ai choisi des standards de métadonnées adaptés au web sémantique comme RDA (Resource, Description & Access), ainsi que IdRef,

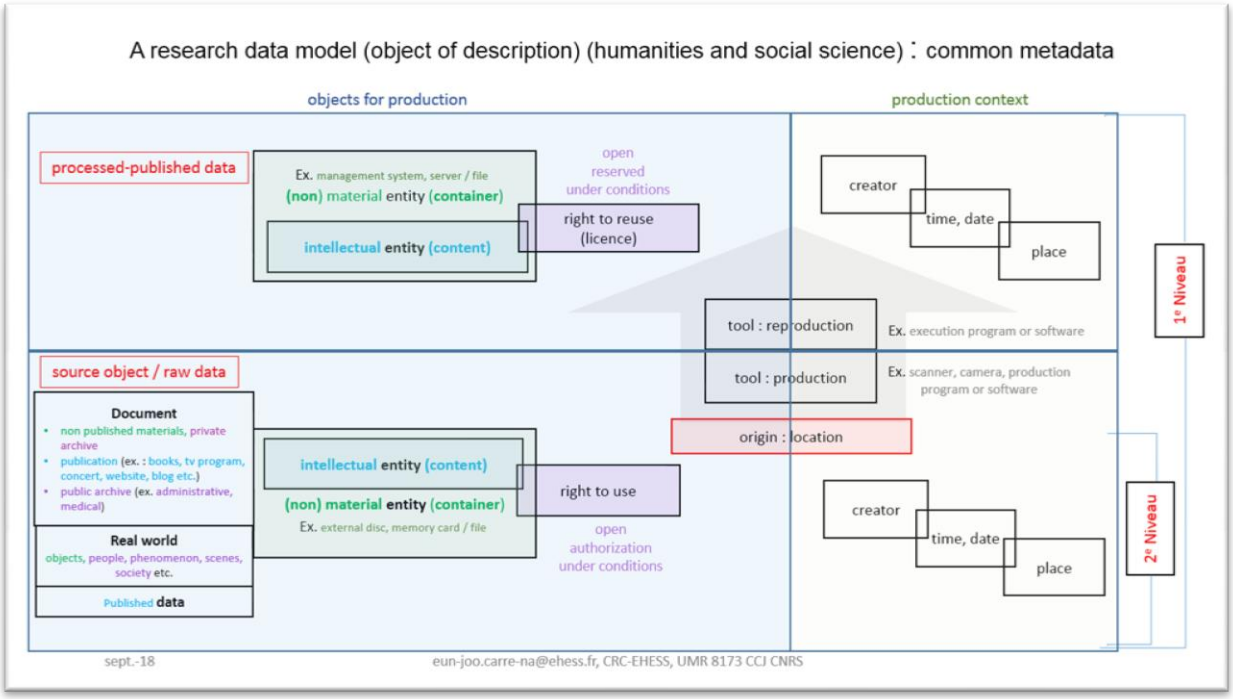
référentiel national qui permettra d'établir des connexions nationales et internationales, par une collaboration avec l'ABES.

3.1. RDA (Ressource : Description et Accès)

RDA, Ressource : Description et Accès, est « une norme de contenu » développée pour suivre l'évolution du web sémantique. Elle comprend les données bibliographiques et d'autorité, points d'accès, et relations ; et elle est conforme aux principes internationaux de catalogage. Elle sera probablement le futur standard international de catalogage.

D'après la « transition bibliographique, 2015 », RDA couvre tous les types de ressources en particulier les ressources électroniques. Elle est souple et extensible pour prendre en compte de nouveaux types de ressources. Son code est ouvert : peu d'éléments obligatoires ; nombreuses options ; totale liberté pour l'encodage des données.

3.2. Modèle de description : champs de métadonnées communes



On peut visualiser ainsi un modèle de données de la recherche comme objet de description. Nous avons déjà remarqué dans le cycle de vie les 2 objets de production : « source » et « produit ». Ils ont respectivement leur propre contexte de production. Chaque objet de donnée contient un « aspect matériel (conteneur) » et un « aspect intellectuel (contenu) ».

Pour les données de la recherche déposées dans le serveur de l'entrepôt (support matériel), ce sont des entités « dématérialisés et virtuelles » sous forme de « fichier informatique (support virtuel) ». Chaque élément composant pourra créer un champ de métadonnée.

Egalement, le droit à gérer aussi varie selon l'étape de production : « droit d'utilisation » vis-à-vis de l'objet-source vs « licence de réutilisation » pour la donnée produite.

3.3. « Données de la recherche » : définition avec RDA

Published data = « computer (media type) » + « online resource (carrier type) » (RDA)

	Carrier type RDA	Media type RDA	File type RDA	Content type RDA
Supports matériels d'images projetées • bobine de film • cartouche de film • cartouche de film fixe • cassette de film • diapositive • film fixe • film fixe court • rouleau de film • transparent pour rétroprojecteur Supports matériels microformes • bande de microfilm • bobine de microfilm • carte à fenêtre • cartouche de microfilm • cassette de microfiches • cassette de microfilm • micro-opaque • microfiche • rouleau de microfilm Supports matériels microscopiques • lame pour microscope Supports matériels sans médiation • feuille • fiche • objet • rouleau • tableau à feuilles mobiles • volume Supports matériels stéréoscopiques • carte stéréoscopique • disque stéréoscopique Supports matériels vidéo • bobine de bande vidéo • cartouche vidéo • cassette vidéo • vidéodisque	Catégorisation qui indique le format du support de stockage et du contenant d'un support matériel combiné avec le type de dispositif de médiation re **équipement ou un système requis** quis pour visionner, faire fonctionner, faire défiler, etc. le contenu d'une manifestation Objet-source **Supports matériels informatiques** • **computer tape cartridge** • **computer card** • **computer chip cartridge** • **computer tape cassette** • **computer disc cartridge** • **computer tape reel** • **computer disc** • **online resource** Supports matériels audio • bobine de bande audio • bobine de fil audio • bobine de piste sonore • cartouche audio • cassette audio • courroie audio • cylindre audio • disque audio • rouleau audio Published data : Material : serveur Virtuel (dematerialized) : fichier numérique	Catégorisation qui indique le type général de dispositif de médiation requis pour visionner, faire fonctionner, faire défiler, etc. le contenu d'une manifestation. • audio • **Computer** • microforme • microscopic projeted • Unmediated • stereoscopic • Video • autre	Type général de contenu des données encodées dans **un fichier informatique**. a) File type : • **audio file** • **data file** • **program file** • **image file** • **text file** • **video file** b) format d'encodage (= MIME type) c) taille du fichier d) résolution e) code de région f) débit binaire codé automatically detectable	Catégorisation qui reflète la forme fondamentale de communication sous laquelle le contenu est exprimé et le sens humain par lequel il est destiné à être perçu. Dans le cas d'un contenu exprimé sous la forme d'une ou plusieurs images, un type de contenu reflète également le nombre de dimensions spatiales au travers desquelles le contenu doit être perçu et la présence ou l'absence perceptibles de mouvement. • cartographic dataset • computer dataset • cartographic tactile three-dimensional form • three-dimensional moving image • tactile three-dimensional form • image animée bidimensionnelle • image animée tridimensionnelle • cartographic image • cartographic moving image • cartographic tactile image • still image • tactile image • notated movement • tactile notated movement • performed music • notated music • tactile notated music • spoken word • computer program • Sounds • text • tactile text

sept.-18 eun-joo.carre-na@ehess.fr, CRC-EHESS, UMR 8173 CCJ CNRS

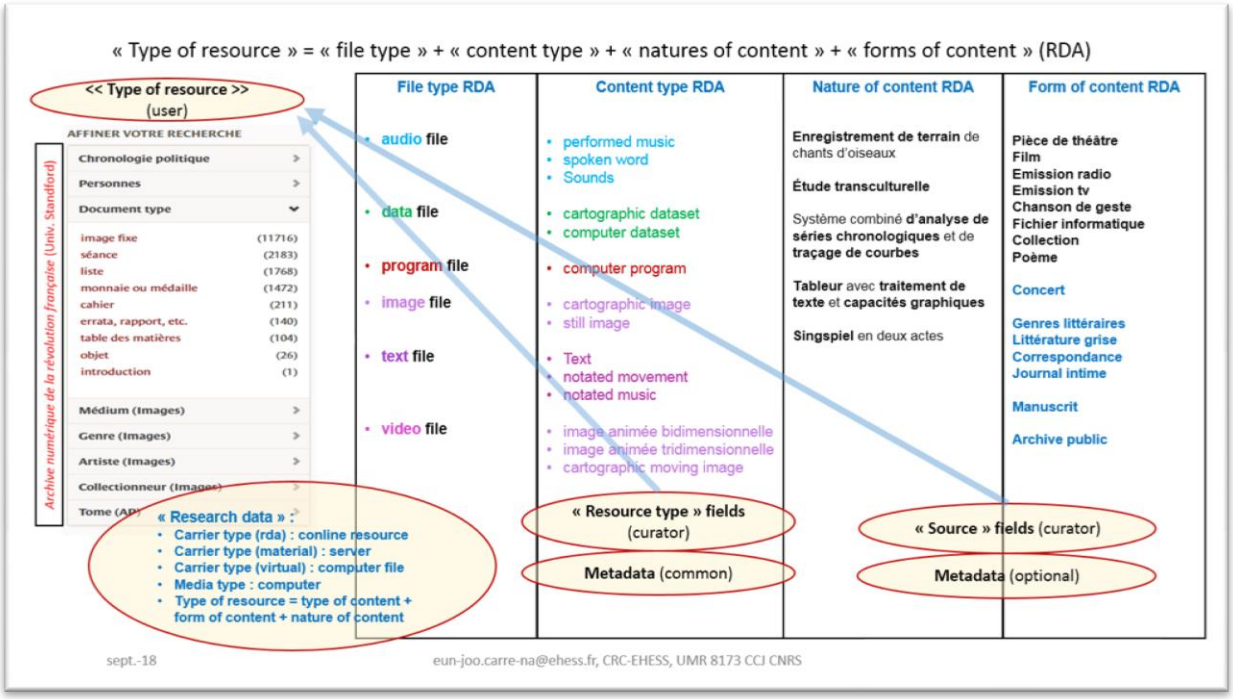
Dans ce tableau, j'ai réuni les différents types et leurs éléments proposés par RDA : « type de support matériel », « type de média », « type de contenu » et « type de fichier informatique ». J'ai marqué en rouge les éléments qui concernent des données de la recherche qui seront déposées dans l'entrepôt.

Si on définit les « données de la recherche déposée dans l'entrepôt » avec les expressions proposées par RDA, elles ont un « type de média informatique » avec un « type de supports matériels informatiques ».

La liste des « supports matériels informatiques » de RDA montre explicitement l'importance de distinction entre « objet-source » et « objet produit » des données de la recherche. A la fin de cette liste, on voit une « ressource en ligne ». Cette « ressource en ligne » concerne nos « données déposées dans l'entrepôt », mais à mon avis, c'est plutôt un intrus dans cette liste : car une « ressource en ligne » elle-même n'est pas un support matériel, mais c'est un élément « dématérialisé » sous forme de « fichier informatique ». Souvent les ressources en ligne sont stockées dans le serveur de leur hébergeur. Donc, pour moi, le « **support matériel** » des « ressources en ligne » est le serveur de l'hébergeur (dans notre cas, le serveur de Didomena), et son « **support virtuel** » est le « fichier informatique » en divers format. Notre affaire de gestion est dans un monde **dématérialisé et virtuel**.

Malgré cette petite confusion dans le type de « support matériel », RDA a bien tenu compte des caractéristiques à décrire du « fichier informatique » : « type de fichier », « format d'encodage », « taille de fichier », etc. La plupart de ces éléments sont détectables automatiquement.

3.4. « Types de ressource » : ensemble des « Types des composants RDA »



Définir des métadonnées sert à une bonne gestion, mais aussi à mieux répondre aux besoins des utilisateurs. Quand on parle de « type de ressources », il est important de distinguer 2 interfaces différentes : celle d'utilisateur et celle de gestionnaire (ou de déposant dans le cas de Didomena).

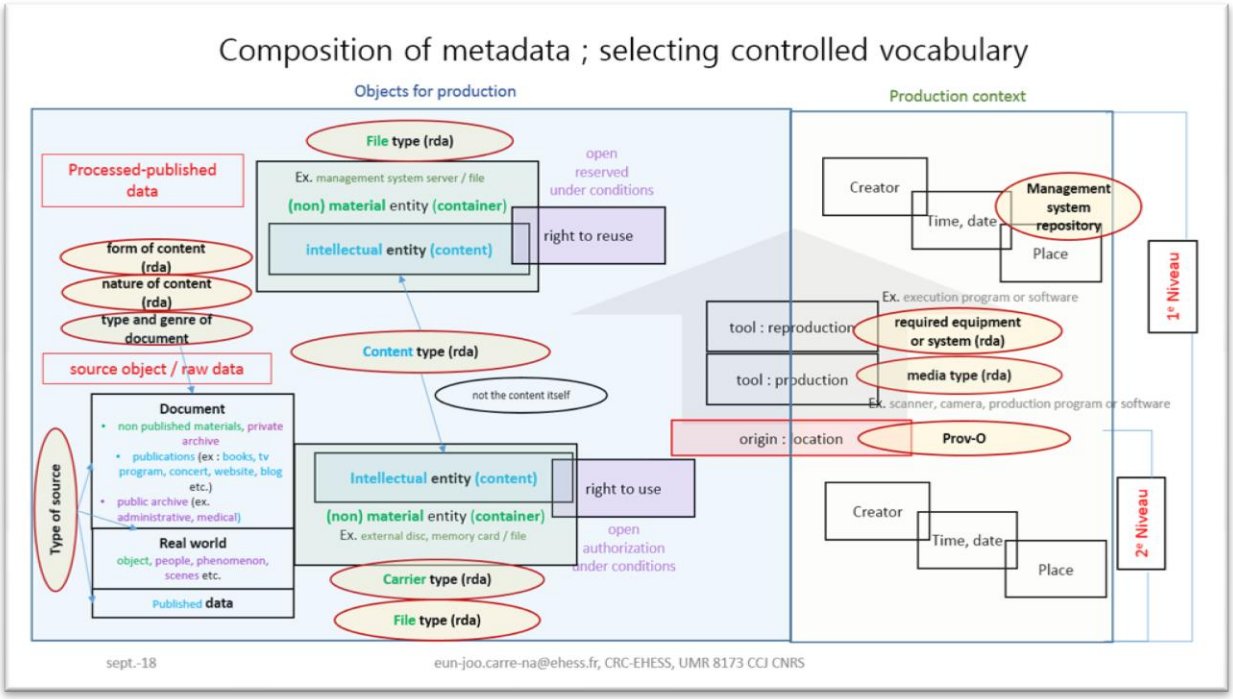
Sur l'interface « **utilisateur** », les « types de ressources » sont affichés souvent sous forme de facette pour faciliter la recherche des utilisateurs. Donc, il faudrait une typologie qui représente les **spécificités** des données du projet concerné. On peut en voir un exemple dans les « archives numériques de la révolution française » de l'université Standford, insérés dans le tableau : « monnaie ou médaille », « cahier », « errata, rapport », « table de matières ». Il s'agit d'une typologie plutôt atypique, mais qui représente bien la spécificité de ses données.

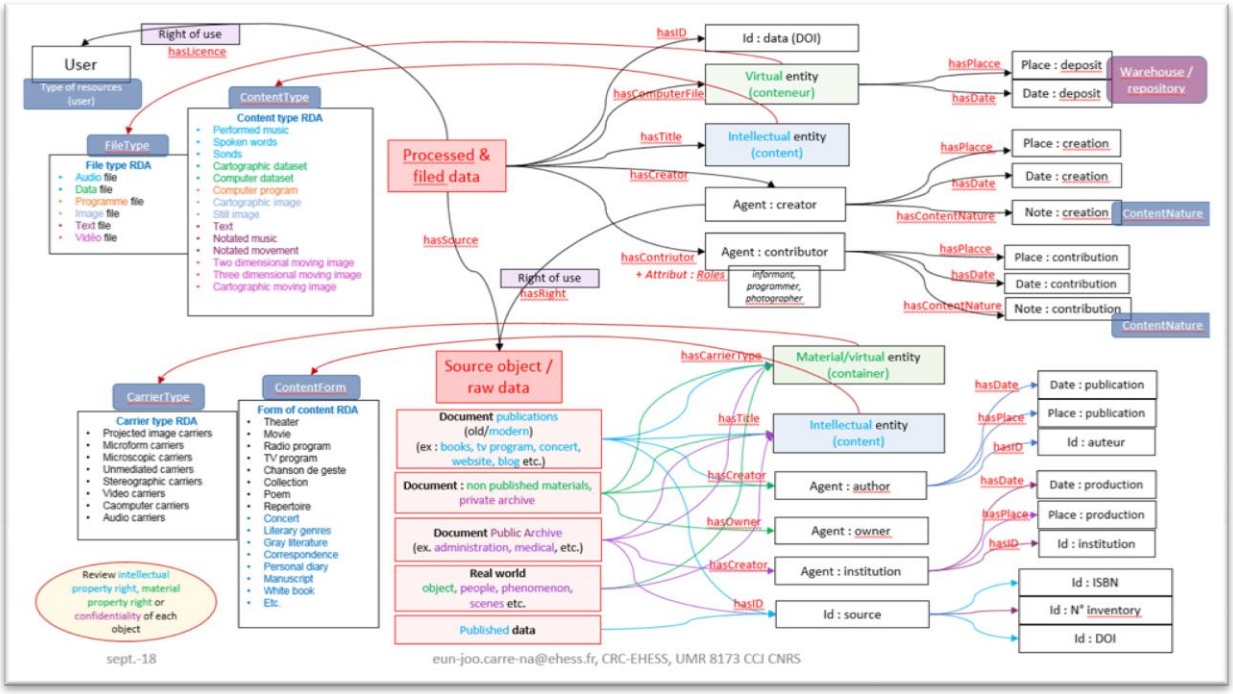
Dans **l'interface de gestion** générale de l'entrepôt qui doit accueillir des données de natures très variées, il faudrait des moyens de décrire tous les types de ressources possible. Mais quelle que soit la typologie définie, elle ne pourra jamais couvrir tous les types de données déposées. Donc, à mon avis, la meilleure solution serait de présenter des éléments-atomes considérés comme composants de la description de divers genres de typologie, avec la possibilité de les combiner librement.

Ainsi, chaque projet pourrait constituer une typologie qui lui convient pour son propre site-web, en combinant des éléments proposés. Les « types de ressources » de **l'interface des utilisateurs** pourraient être personnalisés à travers la combinaison de divers éléments composants de types à choisir sur **l'interface de dépôt**.

RDA a déjà développé certaines parties. A part les « type de support », « type de média » et « type de contenu », on peut compléter la description du contenu avec les champs comme « nature de contenu » et « forme d'une œuvre (contenu) ». Surtout la « forme d'une œuvre (contenu) » est assez libre pour qu'on puisse mettre différents types de documents et genres littéraires. On pourra utiliser ce champ pour les « notes de production » comme « cahier de terrain ».

Voici le modèle de description avec ses vocabulaires contrôlés possibles de RDA à chaque champ de métadonnée.





3.5. Limite de RDA et autres métadonnées

Etant donné que RDA est en cours de développement, c'est un code encore incomplet pour certains types de ressources : audiovisuelles, continues, cartographiques, images fixes. Pour autant, comme RDA est un code ouvert et très libre, on peut le combiner avec d'autres vocabulaires contrôlés (comme Prov-O, CDAT, METS) pour des métadonnées spécifiques et optionnelles. Par exemple, pour

les données audiovisuelles, on peut utiliser Prov-O, ce qui est recommandé par le ministère de la culture pour tracer les sources, afin de gérer le droit d'auteur.

3.6. Référentiels : choix du réseau professionnel (IdRef- ABES)

Didomena a 3 référentiels : Vivo, création des déposants, et IdRef. Etant donné que IdRef est un référentiel national développé par l'ABES pour suivre l'évolution du web sémantique, cela permettra à Didomena d'établir des connexions au niveau national voir international.

Au lieu de développer mon explication sur IdRef, je vous invite à la lecture du support de formation de François Mistral (URFIST-Paris, 2018). Je résume seulement les 3 points essentiels sur l'importance et l'utilité du point d'accès d'IdRef.

1) La connexion avec l'IdRef va mettre Didomena en relation avec le monde extérieur, voire l'international, et aussi les catalogues Sudoc et BNF, ainsi que diverses Open éditions comme HAL, Persse, etc.
2) Un autre avantage d'IdRef est la facilité de la gestion des points d'accès. On peut créer et modifier directement sur l'interface-client, le site-web d'IdRef. Dans l'interface de gestion du site-web d'IdRef, chaque champ est accompagné de l'explication, on peut remplir des champs sans avoir la connaissance sur la norme UNIMARC.
3) Le point le plus fort d'IdRef est de réunir plusieurs identifiants comme SINI, les identifiants de BNF, de HAL, mais aussi ORCID.

3.7. Didomena : sa configuration et ses caractéristiques

Dans les sciences de la nature, des entrepôts de données de la recherche sont souvent conçus par domaine de recherche, comme « open data de l'agriculture », « data center de la biologie », etc. Mais d'après l'observation de quelques projets - pilote, Didomena, en tant qu'un entrepôt institutionnel en SHS, va accueillir des données de natures très variées. Pour décrire la configuration de Didomena et les caractéristiques des diverses données déposées, j'ai utilisé « 3 types de numérisation » définis selon le « degré de numérisation » à la rencontre annuelle du réseau DocAsie en 2016. Faute de découverte de noms convenables, je nomme respectivement ces 3 types avec leur exemple-type : « bibliothèque numérique », « édition numérique » et « donnée élaborée ».

Les critères utilisés en 2016 étaient « objectif », « outil » et « format ».

Type de numérisation	Outils de travail
Image fixe	Photographie, scanner (jpeg, png, gif, etc.)
Rérérencements	Méta-données, SIGB, OCR, saisie manuelle de text (pdf, word, dublin core, rdf, etc.)
Applications numériques	Encodages (XML, TEI, etc.) éditeurs XML (oXygen, etc.)

Cette fois, en ajoutant le classement des types de RDA parmi les critères, j'ai pu relever différents degrés de modification entre l'objet-source / donnée brute et la donnée élaborée-publiée, comme le montre le tableau ci-dessous. Cette différence de degré de modification entre la source et le produit peut faire sortir plus clairement les caractéristiques des types des données déposées.

Configuration and features

Type of digitization	Objective	Content (work)	Type of carrier	Type of media	Type of content	Form of content	Examples
1	Conservation	Unchanged	Changed	Changed	Unchanged	Unchanged	Image digitization (ex. digital library, museum)
2	Reference research	Unchanged	Changed	Unchanged or changed	Unchanged	Unchanged or changed	Scanning with intervention on the content either by encoding (xml) or by OCR (ex. digital edition)
3	Scientific purpose	Changed	Changed	Changed	Changed	Changed	Production of new data (ex. statistics, corpus, geographic data)

Characteristics of the "research data" defined according to the degree of modification between the source / raw data and the elaborate-published data, according to "3 types of digitaliization" (Eun-joo Carré-Na, 2016)

sept.-18 eun-joo.carre-na@ehess.fr, CRC-EHESS, UMR 8173 CCJ CNRS

Le 1er type est un exemple du type « bibliothèque numérique ». Les données de ce type ont souvent des documents non-numériques comme source. Leurs documents-source sont numérisés avec l'objectif, entre autre, de garder l'état originel. Cette numérisation sert pour la mission de conservation du patrimoine culturel par la bibliothèque, ainsi que pour fournir des preuves factuelles dans le domaine de la recherche scientifique. Donc, à part le « type de support matériel », il y a peu de changement, surtout au niveau du contenu.

Le 2e type est représenté par un exemple d'« édition numérique ». Ce type de donnée peut être issu soit d'une numérisation, soit de la production numérique. Ces moyens de production font varier le « type de média » et la « forme de son contenu ». L'objectif de cette production, qu'elle soit numérisée ou originellement numérique, est de rendre accessible par la machine non seulement les paramètres de métadonnée du document concerné, mais aussi son contenu. Le « contenu » mis en ligne comme référence trouvable par le moteur de recherche peut être repéré soit comme des signes ou caractères, soit comme un contenu sémantique. Cela dépend de l'encodage sur le contenu. Dans ce type de donnée, les « média », « forme » et « support » peuvent changer mais l'ensemble du contenu sémantique lui-même ne change pas.

Ces 2 types de données peuvent être produits sans être forcément liées à un projet scientifique, mais avec le simple objectif de valoriser le document en question. Pour autant, ils servent souvent à justifier des arguments ou résultats scientifiques en SHS en tant que « données de la recherche », ou à valoriser le résultat d'un projet de recherche.

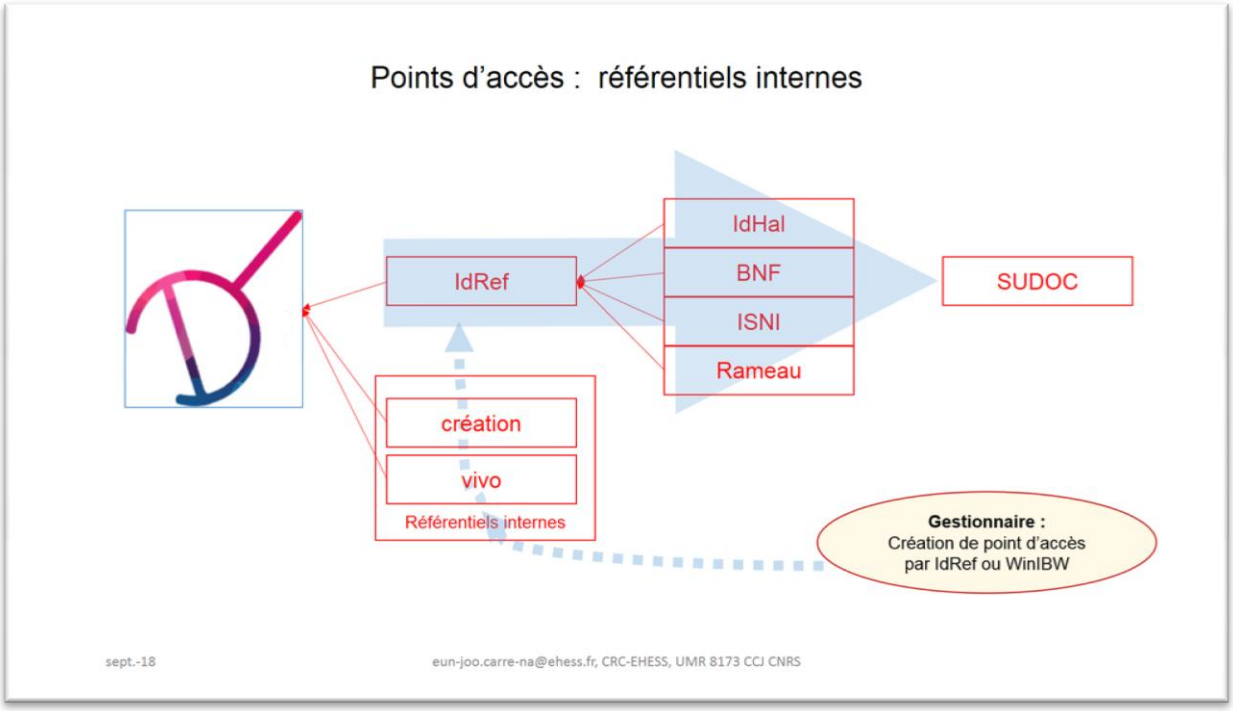
Le 3e type est représenté par des « données élaborée-créée ». Il s'agit d'une création d'une nouvelle œuvre, bien qu'il ne change pas le principe « factuel » de sa source. Un chercheur peut créer une nouvelle donnée sous forme de statistique ou d'un tableau évolutif, à partir d'une masse d'éléments factuels (documents ou phénomènes) collectées.

4. Evolutions souhaitées pour une bonne gestion

Avec le choix de RDA et IdRef, il est certain que Didomena est sur une bonne voie pour ses éventuelles évolutions administratives et techniques.

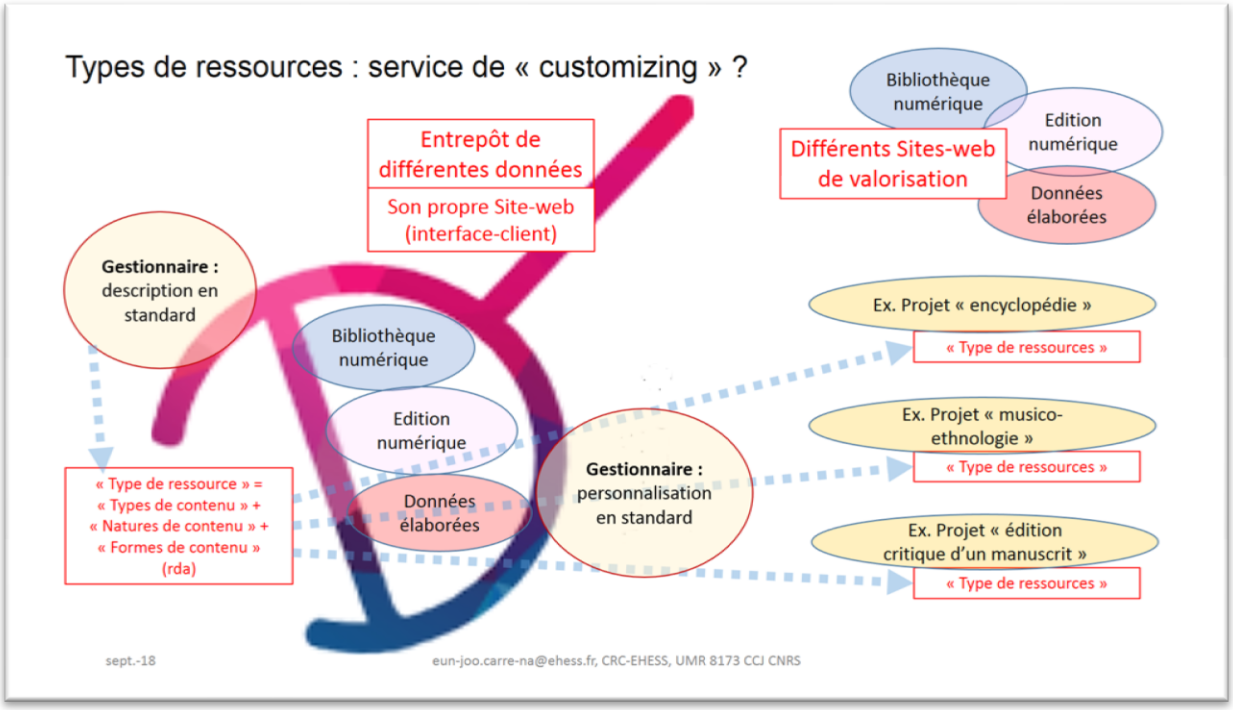
Puisque Didomena est en pleine phase de développement et de conception, son évolution reste ouverte, et il serait hasardeux de préjuger de son fonctionnement sans une excellente compréhension de l'état actuel de ses aspects techniques que je ne domine pas ; j'évoquerais cependant quelques évolutions souhaitables en son état actuel sur le plan de la gestion, afin de suggérer quelques services possibles de la part des gestionnaires.

4.1. Gestion des référentiels et points d'accès



Tandis que le réseau IdRef emmène Didomena vers diverses connexions externes comme des catalogues nationaux et internationaux via différents points d'accès (nouveau terme par la transition bibliographique pour dire « autorité »), les référentiels internes de Didomena, via Vivo ou la création des déposants n'ont pas ces points d'accès. Les gestionnaires peuvent apporter leur contribution, en créant et gérant des points d'accès à travers de l'IdRef.

4.2. Description en élément-atome à combiner pour le service de « customizing »



Didomena est équipé de champs de métadonnée standardisés en RDA. Pourtant, le fait que Didomena a choisi de laisser la gestion aux déposants de chaque projet, impose une forte subjectivité dans les interprétations et des risques d'incompréhension, ou de contresens de ces champs de métadonnée, ou bien une négligence dans les procédures de dépôt. Ces risques pourraient être diminués par un service des gestionnaires de la description en standard dans un espace de gestion (étape de curation) mis en option.

4.3. Aide à la personnalisation de typologie

Cela sera pareil pour les champs de divers types. Puisque Didomena a choisi la politique de « non-gestion » des projets et que la gestion des données sera à la charge de déposants qui ne sont pas toujours des gestionnaires professionnels, Didomena souhaite que les déposants ne soient pas effrayés par la longueur de listes d'éléments de type à choisir. Pour cette raison Didomena a choisi le « confort des déposants ». Donc, en son état actuel de développement, Didomena ne propose que des champs des types de ressources, évolués à partir des « types de contenu » RDA avec quelques ajouts de certains types, estimés utiles aux utilisateurs.

Les champs complémentaires comme « nature » et « forme » du contenu ne sont pas encore appliqués. Avec ces champs, on peut imaginer un « service de personnalisation ». Les gestionnaires pourraient aider chaque projet dans le classement des types, afin de créer un « type de ressources » qui correspond le mieux aux objectifs du projet en question, pour représenter au mieux sa spécificité. Ou bien, en laissant un maximum d'éléments-atome qui définissent différentes typologies pour que les déposant puissent choisir, on pourrait faire générer (semi-)automatiquement « le type de ressources » le plus adapté à chaque projet.

Bibliographie (l'ordre de citation)

- OCDE, *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, 2007, Paris, http://www.oecd.org/fr/science/sci-tech/38500823.pdf
- Pierre-Yves Arnould, Marie-Christine Jacquemot, *Guide de bonnes pratiques Gestion et valorisation des données de la recherche*, 2016, https://hal.inria.fr/hal-01275841/
- Joachim Schöpfel, Eric Kergosien, Hélène Prost, « « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse », *Atelier VADOR : Valorisation et Analyse des Données de la Recherche; INFORSID 2017*, May 2017, Toulouse, France. 2017, https://hal.univ-lille3.fr/hal-01530937
- Sylvie Cocaud, « Les entrepôts de données de recherche », *Participer à l'organisation du management des données de la recherche, gestion de contenu et documentation des données*, Action Nationale de Formation RENATIS – MEDICI 2017, https://anfdonnees2017.sciencesconf.org/data/pages/Entrepots_ANFRenatis_2017_Cocaud_Aventurier_1.pdf
- Pain M. (2016). *Les données de la recherche et leurs entrepôts, de la documentation à la réutilisation: étude de cas pour l'archive HAL*. Mémoire de Master. Enssib, Villeurbanne. https://memsic.ccsd.cnrs.fr/mem_01374509
- Bibliothèque de l'université Standford, *Archive numérique de la révolution française,* https://frda.stanford.edu/fr
- Eun-joo Carré-Na, « Réflexions sur les rôles des documentalistes dans la numérisation : le cas du « Répertoire historique de l'administration coréenne » de Maurice Courant », *10 e Rencontre annuelle du réseau DocAsie : les fonds asiatiques à l'ère du numérique*, Paris, 2016, https://hal.archives-ouvertes.fr/hal-01337488
- Ph. Le Pape, *Que cent zones zéro s'épanouissent*, ABES, 2014, https://rda.abes.fr/2014/05/28/que-cent-zones-zero-sepanouissent/
- Ph. Le Pape*, La zone zéro*, ABES, 2014, https://rda.abes.fr/2014/05/26/la-zone-zero/
- Ph. Le Pape, *RDA : à quoi sert l'élément Type de support ?,* ABES, 2017
- Transition bibliographique**, «** 4. RDA : Resource Description and Access », *Sensibilisation à l'évolution des catalogues,* RNF, 2015, https://www.transition-bibliographique.fr/wp-content/uploads/2016/01/support_formation_rda.pdf
- Transition bibliographique**,** « Type de médiation, Type de contenu », *Application des règles de catalogage RDA-FR*, 2017, https://www.transition-bibliographique.fr/wp-content/uploads/2017/04/RDA_FR2016_SupportFormation_Type_m%C3%A9diation_contenu.pdf
- Transition bibliographique**, «** III. Description des supports, matériels et des contenus », *Application des règles de catalogage RDA-FR*, 2017, https://www.transition-bibliographique.fr/wp-content/uploads/2017/04/RDA_FR2017_SupportFormation_TypedeSupport.pdf
- François Mistral, *IdRef – Identifiants et Référentiels pour l'Enseignement supérieur et la Recherche : interopérabilité et visibilité sur le web*, URFIST-Paris, 2018