Le Livre Blanc du consortium Mémoires des Archéologues et des Sites Archéologiques : Guide des bonnes pratiques numériques en archéologie

Olivier Marlet, Bruno Baudoin, Loup Bernard, Laure Bézard, Romain Boissat, Pierre-Yves Buard, Agnieszka Halczuk, Florian Hivert, Jamet Hélène, Blandine Nouvel, et al.

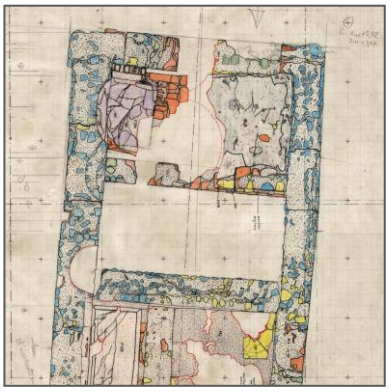
Le **Livre Blanc**

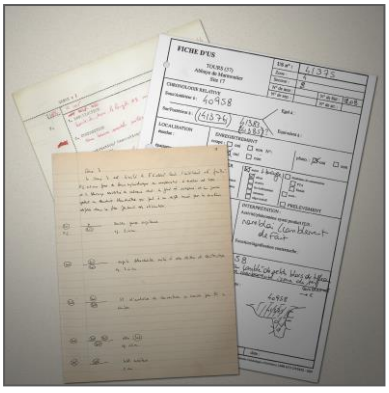
du consortium Huma-Num

Mémoires des Archéologues et des Sites Archéologiques

Guide des bonnes pratiques numériques en archéologie

Version 0.1 (décembre2021)





Auteurs

Olivier Marlet (dir.), CNRS, UMR 7324 CITERES-LAT, Tours

Bruno Baudoin, CNRS, UMR 7299 Centre Camille Jullian, CNRS, Aix Marseille Université, Aix-en-Provence

Loup Bernard, Université de Strasbourg, UMR 7044 Archimède

Laure Bézard, CNRS, Maison de l'Orient et de la Méditerranée – Jean Pouilloux, Lyon

Romain Boissat, CNRS, Maison de l'Orient et de la Méditerranée – Jean Pouilloux, Lyon

Pierre-Yves Buard, Université de Caen Normandie, Pôle Document Numérique, USR 3486 Maison de la Recherche en Sciences Humaines

Agnieszka Halczuk, Prestataire, Maison de l'Orient et de la Méditerranée – Jean Pouilloux, Lyon

Florian Hivert, contractuel MASA, USR 3501 Maison des Sciences de l'Homme Val de Loire, Tours

Hélène Jamet, CNRS Maison de l'Orient et de la Méditerranée – Jean Pouilloux, Lyon

Blandine Nouvel, CNRS, UMR 7299 Centre Camille Jullian, Aix Marseille Université, Aix-en-Provence

Xavier Rodier, CNRS, UMR 7324 CITERES-LAT, USR 3501 MSH Val de Loire, Tours

Miled Rousset, CNRS, Maison de l'Orient et de la Méditerranée Jean Pouilloux, Lyon

Lizzie Scholtus, contractuelle MASA, USR 3501 Maison des Sciences de l'Homme Val de Loire, Tours

Institution émettrice : Consortium MASA

contact mail du projet : olivier.marlet@univ-tours.fr

Sommaire

I. Préambule : Le consortium MASA

Le consortium Mémoires des archéologues et des sites archéologiques (MASA), de la très grande infrastructure de recherche IR* Huma-Num, est né à la fin de l'année 2012 de l'expérience acquise par et au sein de plusieurs Maisons des Sciences de l'Homme dans le domaine de l'archéologie et du traitement de la documentation produite par les archéologues (Fig. 1).

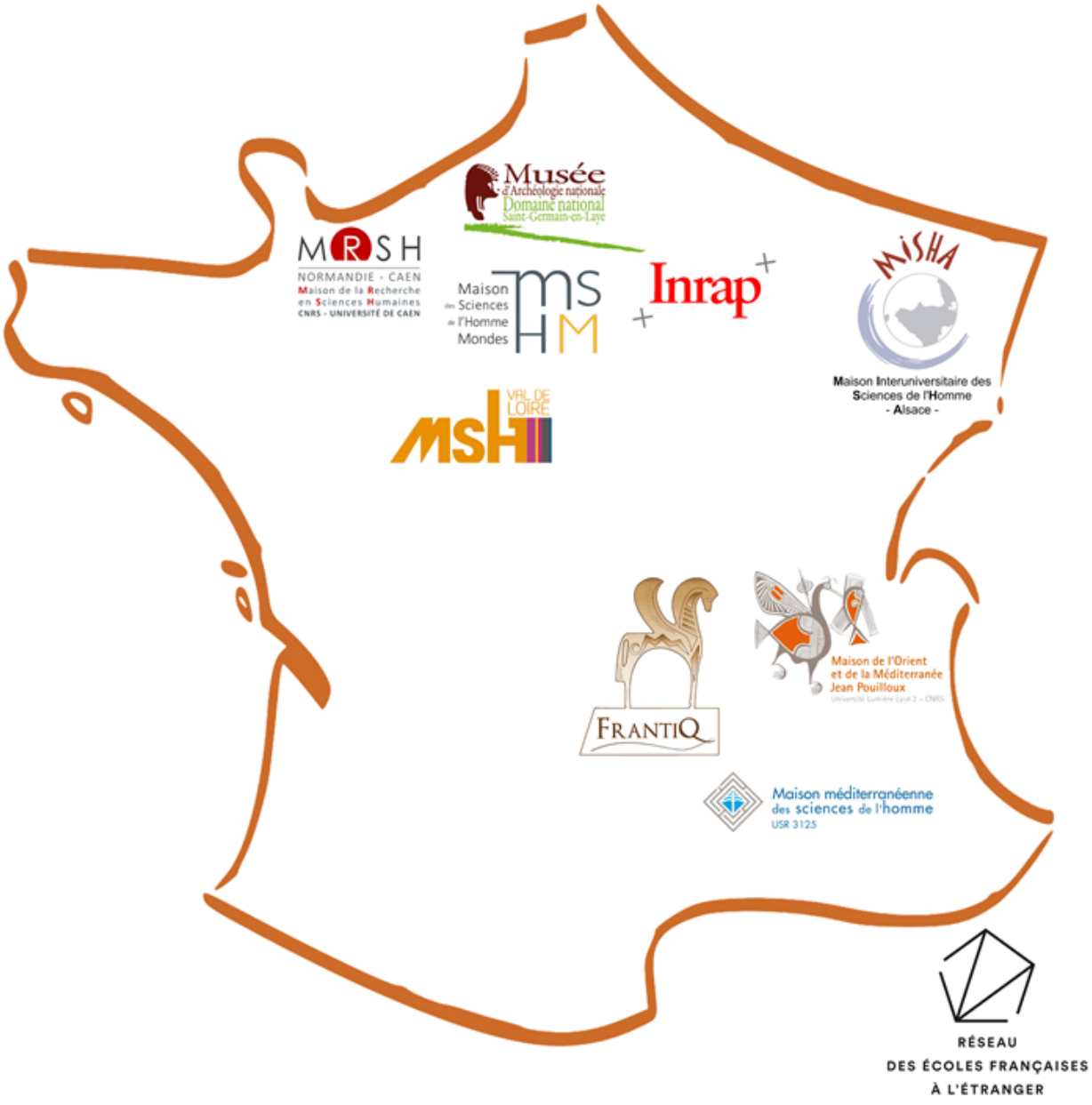


Fig. 1 : Les acteurs du consortium MASA

Fondé sur le constat de la nécessité de préserver la masse considérable de données accumulées par les archéologues, il s'est appuyé sur les compétences de ses partenaires pour faire la démonstration du traitement, de la numérisation à la publication, de corpus d'archives et de données archéologiques. Les résultats proposés ne sont pas les seuls à aller dans le sens du partage et de la publication numérique des données. Ils ne couvrent pas tous les aspects de l'archéologie et ne sont pas les uniques exemples à suivre. En revanche, ils sont conçus comme des cas d'école qui permettent d'apporter des réponses aux besoins de la communauté archéologique comme autant de preuves de concept.

Renouvelé en 2017, les efforts du consortium ont été concentrés sur la pérennisation et la mise à disposition d'archives

au format numérique, sur la réutilisation de corpus constitués, sur l'interopérabilité des systèmes d'information, sur le partage numérique des données archéologiques. Afin de répondre aux besoins identifiés de la communauté archéologique et de faire face à l'urgence de sauvegarder les corpus d'archives et les bases de données existantes, le consortium MASA a élaboré un écosystème numérique fondé sur le cycle de vie des données et centré sur la mise en œuvre des principes FAIR, s'inscrivant ainsi résolument dans la dynamique de la science ouverte. Pour atteindre ses objectifs, le consortium a constitué son écosystème à partir d'outils open source et de référentiels et standards internationaux lorsqu'ils existaient ou, dans le cas contraire, à les développer (Plan de Gestion de Données spécifique à l'archéologie, OpenGuide, OpenArchaeo, LogicistWriter) ou à soutenir leur développement (Opentheso, PACTOLS, ArkeoGIS). Cette démarche a conduit MASA à s'insérer dans la communauté internationale pour y partager les bonnes pratiques, les outils et les jeux de données à travers des projets tel que le H2020 ARIADNEplus qui rassemble plus de 40 partenaires en archéologie et en informatique, mais également sur la diffusion des principes FAIR (H2020 PARTHENOS, COST SEADDA), le développement de l'ontologie du CIDOC CRM (CRM SIG) et des outils pour le manipuler (3M, Ontop). Au-delà de la communauté archéologique, les travaux conduits en application des principes FAIR ont amené à des collaborations dans le vaste domaine du patrimoine culturel, avec le consortium 3D-SHS pour l'interopérabilité des données, avec le musée du Louvre, la BNF et les Archives nationales sur le Web sémantique, ou encore avec le Chantier scientifique Notre-Dame pour la gestion des données numériques et leur interopérabilité. Cet élargissement thématique a également lieu à l'échelle internationale avec la mobilisation de l'expertise du consortium sur les données du patrimoine culturel dans le projet de préfiguration d'un Centre de compétences européen pour la conservation du patrimoine culturel (4CH - H2020). À l'issue de cette seconde période de quatre ans, les points forts de MASA sont la FAIRisation des données qui est largement engagée, des preuves de concepts opérationnelles et le développement de l'ancrage dans la communauté internationale.

Labellisé à nouveau pour les deux dernières années de cette première phase, MASA est un vecteur de diffusion des principes FAIR qui s'est fixé comme objectif la mobilisation large de la communauté archéologique française afin d'accompagner son engagement massif dans la science ouverte. Tout en poursuivant le développement des projets engagés, ces deux années sont consacrées au passage à l'échelle tant pour les corpus de données que pour les services offerts.

Fort de compétences variées mises en commun durant une décennie au sein du consortium et en partenariat avec les infrastructures nationales et internationales, MASA pense avoir atteint une maturité suffisante pour vous proposer cette première version de sa compilation de ressources structurées autour du cycle de vie des données. Conscient de ne pas maîtriser de manière équivalente tous les aspects de cycle de vie, les membres du consortium MASA, auteurs de ce Livre Blanc, ont tenté *a minima* de fournir des pistes pour aider les chercheurs dans leur démarche de partager leurs données. L'ambition de ce Livre Blanc est d'accompagner les archéologues dans cette optique de partage de données afin de contribuer à l'émergence des données archéologiques au sein du *Linked Open Data*.

II. Objectifs du Livre Blanc

Les objectifs de ce Livre Blanc sont de synthétiser l'ensemble des bonnes pratiques numériques tout au long du cycle de vie des données pour accompagner la communauté archéologique à publier ses données dans le Web sémantique en assurant leur diffusion.

1. Évaluer la qualité des jeux de données numériques

Les bonnes pratiques pour publier de l'information numérique de qualité ne sont pas propres à une discipline en particulier, mais, dans le détail, les approches de chaque discipline peuvent varier sensiblement. Ce guide propose une approche adaptée à l'archéologie. La première étape pour juger de la qualité d'un jeu de données numérique est de procéder à son évaluation. Plusieurs outils sont disponibles, correspondant à deux méthodes éprouvées : les « *Five Stars Linked Open Data* » et les « *FAIR principles* ».

Ressources:

- Programme *e-Learning* du Portail Européen de Données : https://data.europa.eu/elearning/fr
- Le portail *Archaeological Data Service* (ADS) au Royaume Uni met à disposition plusieurs guides de bonnes pratiques régulièrement mis à jour et enrichis.
- *Research Data Alliance* (RDA) : catalogues des recommandations à l'échelle internationale.

A. Les 5 étoiles du Linked Open Data

En 2010, Tim Berners-Lee, principal inventeur du *World Wide Web*, a proposé une méthode basée sur 5 étoiles pour évaluer le niveau d'ouverture des données sur le *Linked Open Data* (Fig. 2).

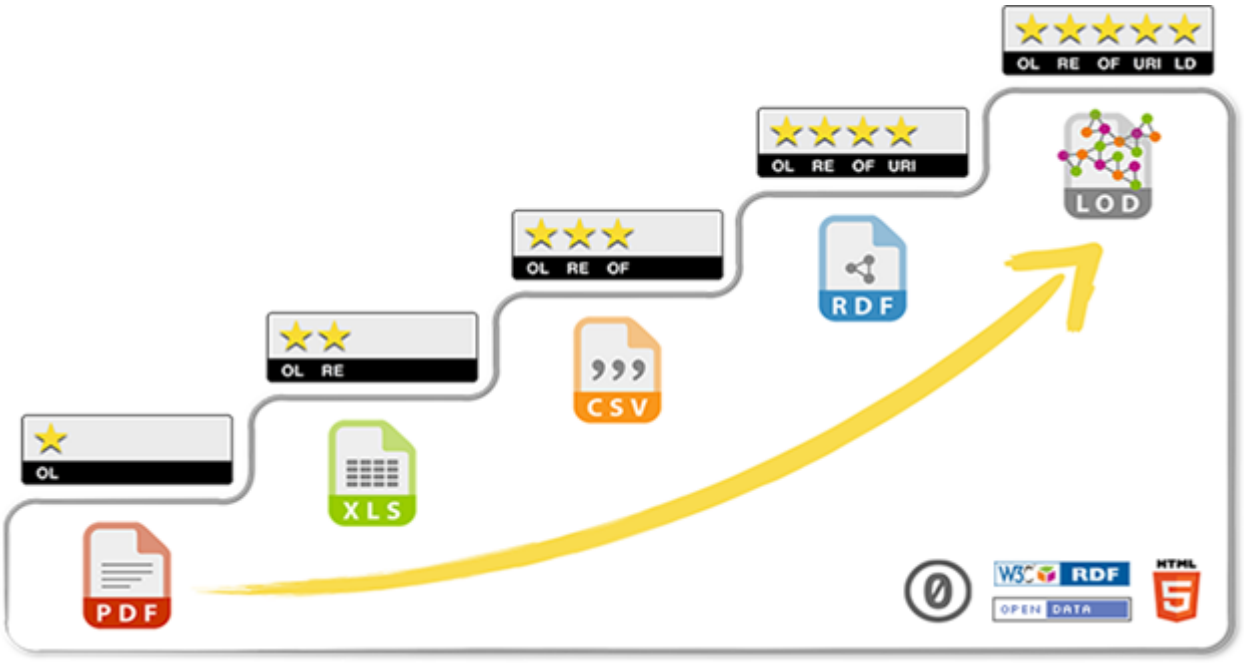


Fig. 2 : Le système de notation à 5 étoiles du Linked Open Data

- Première étoile (*On Line*) : les données sont en ligne, accessibles (licence ouverte) et lisibles. C'est le cas d'un fichier PDF, que ce soit du texte ou une image issue d'un texte scanné.
- Deuxième étoile (*machine Readable*) : les données sont dans un format structuré et donc lisibles par une machine et exportables vers d'autres formats structurés. C'est le cas d'un fichier MS Excel.
- Troisième étoile (Open Format) : les données sont dans un format ouvert et non propriétaire. C'est le cas d'un fichier CSV qui n'est la propriété d'aucun éditeur.
- Quatrième étoile (Uniform Resource Identifier) : les données sont référencées à l'aide d'URI, c'est-à-dire identifiées de manière unique sur le réseau (une adresse web par exemple). C'est le cas d'un fichier RDF composé de triplets. Cette étape nécessite des compétences complémentaires pour la manipulation des graphes RDF, différente de celle de données tabulaires ou d'une arborescence XML.
- Cinquième étoile (Linked Data) : les données sont liées à des référentiels du Linked Open Data (thésaurus normalisé, Wikidata, etc.). Cette dernière étape est moins technique que la précédente mais nécessite néanmoins d'aligner les concepts du jeu de données avec des référentiels partagés. Plus le référentiel est partagé par la communauté, plus l'interopérabilité des données est efficace.

Ressources :

- Il existe des outils en ligne pour évaluer la qualité des données au regard des *5 stars LOD*, comme l'*Open Data Certificate* : https://certificates.theodi.org/en/.

Une équipe de chercheurs de l'*Aalto University* en Finlande a complété ces 5 étoiles par 2 supplémentaires : la sixième étoile est accordée si un modèle de la structuration des données est fourni ; la septième étoile est accordée si la structure des données publiées respecte ce schéma.

Ressources :

- Hyvönen, Tuominen, Alonen and Mäkelä 2014: Linked Data Finland: A 7-star Model and Platform forPublishing and Re-using Linked Datasets, European Semantic Web Conference, DOI:10.1007/978-3-319-11955-7_24

B. Les principes FAIR

En 2016, le groupe FORCE11 publie FAIR Guiding Principles for scientific data management and stewardship dans la revue Scientific Data. Les auteurs y proposent un guide pour faire en sorte que les ressources numériques soient *Findable, Accessible, Interoperable, Re-usable*(en français Faciles à trouver, Accessibles, Interopérables, Réutilisables). L'objectif est de faciliter le traitement des données par les systèmes informatiques pour aider les humains à mieux les gérer, leur volume étant en croissance constante (Fig. 3).

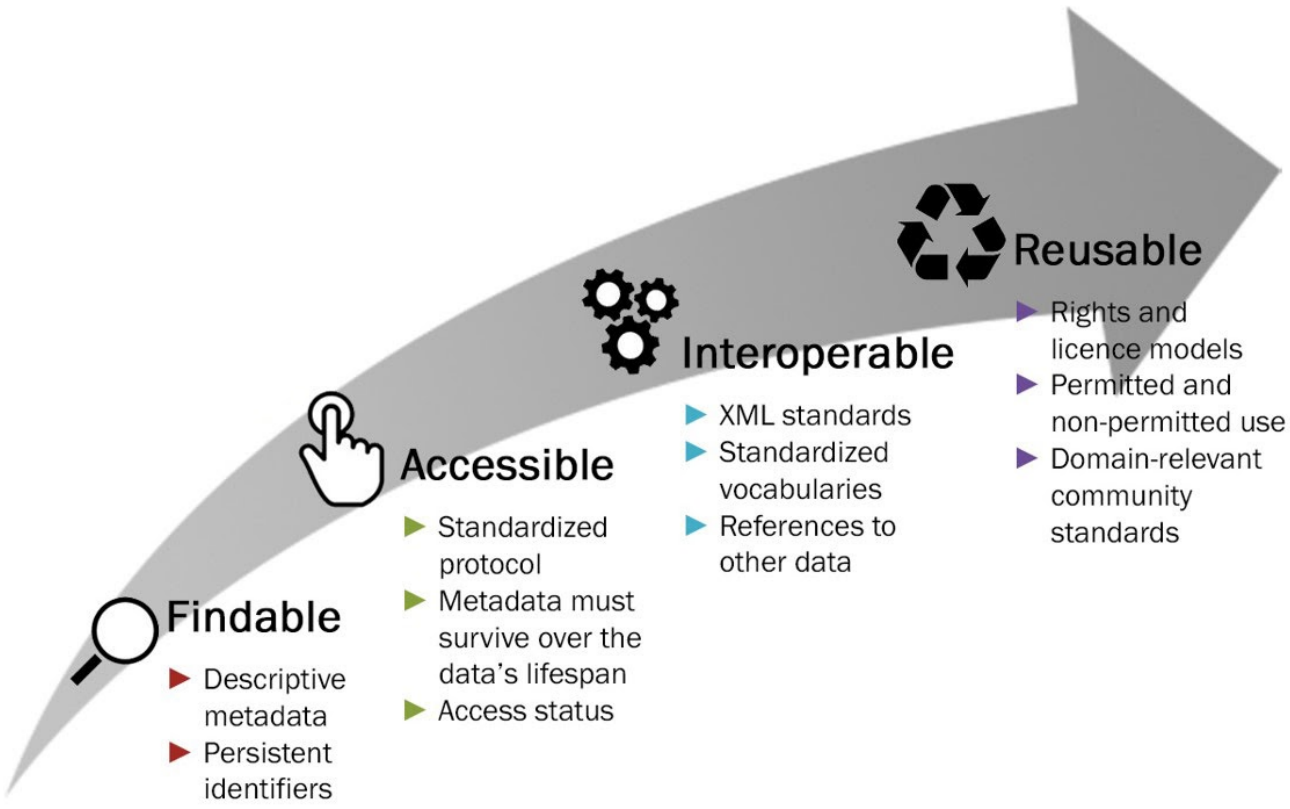


Fig. 3 : Synthèse des principes FAIR

Ressources :

- GO-FAIR : https://www.go-fair.org/fair-principles/
- PARTHENOS : présentation orientée pour la Patrimoine Culturel et accès au guide en français (PDF).
- DORANUM: https://doranum.fr/enjeux-benefices/principes-fair/
- Ouvrir la Science : https://www.ouvrirlascience.fr/fair-principles/
- L'exemple d'application des principes FAIR pour ADS au Royaume Uni.

Il existe des outils en ligne pour évaluer la qualité des données au regard des principes FAIR, ou selon le néologisme

approprié leur degré de FAIRisation :

- le prototype du *Data Archiving and Networked Services* : https://www.surveymonkey.com/r/fairdat, traduit en français par DoRANum : FAIR-Aware.
- L'outil d'une équipe espagnole pour évaluer la FAIRisation d'une ontologie : https://foops.linkeddata.es/FAIR_validator.html.
- La *Research Data Alliance* a également publié une analyse des outils d'auto-évaluation du respect des principes FAIR : https://zenodo.org/record/3629618#.YXqgG55BzIV.

2. Intégration au Standardization Survival Kit (SSK)

Le SSK a été initié dans le cadre du programme H2020 PARTHENOS puis repris par Huma-Num. Il est intégré aux services du Social Sciences & Humanities Open Marketplace de l'ERIC DARIAH-EU. Son objectif est d'aider les chercheurs en stabilisant les connaissances sur les normes et les bonnes pratiques du numérique en SHS. L'organisation du SSK se fait par scénarios avec un objectif et des étapes à suivre pour l'atteindre, ces étapes pouvant être par exemple un tutoriel pour l'utilisation d'un logiciel, l'utilisation d'un service web, etc. Chaque scénario peut être réutilisé dans le cadre d'un autre scénario et être adapté pour les besoins spécifiques d'une discipline donnée.

Le consortium MASA a prévu d'adapter le contenu de ce Livre Blanc en scénarios du SSK pour fournir un accès à la fois plus interactif et pérenne aux bonnes pratiques proposées à la communauté des archéologues et au-delà.

Ressources :

- Standardization Survival Kit (Huma-Num) : http://ssk.huma-num.fr/#/

3. L'écosystème numérique MASA fondé sur le cycle de vie des données

La gestion et le partage des données de la recherche sont au cœur des feuilles de routes des acteurs de la recherche (feuille de route du CNRS pour la Science Ouverte, par exemple), qui s'inscrivent dans les plans nationaux successifs pour la Science Ouverte de 2018 et 2021, prolongés par les 15 feuilles de routes ministérielles sur la politique de la donnée, des algorithmes et des codes sources. Ces recommandations ont pour objectif de garantir la traçabilité et l'intégrité des données produites afin d'en améliorer la conservation et d'en faciliter l'accès, l'interopérabilité et la réutilisation. Le Livre Blanc, MASA est organisé selon le cycle de vie des données de la recherche, cercle vertueux décrivant le processus d'utilisation des données depuis leur création jusqu'à leur réutilisation (Fig. 4). Au modèle de référence élaboré par *UK Data Archive* (Research Data Lifecycle), a été ajouté la planification en amont de la création (Fig.1).

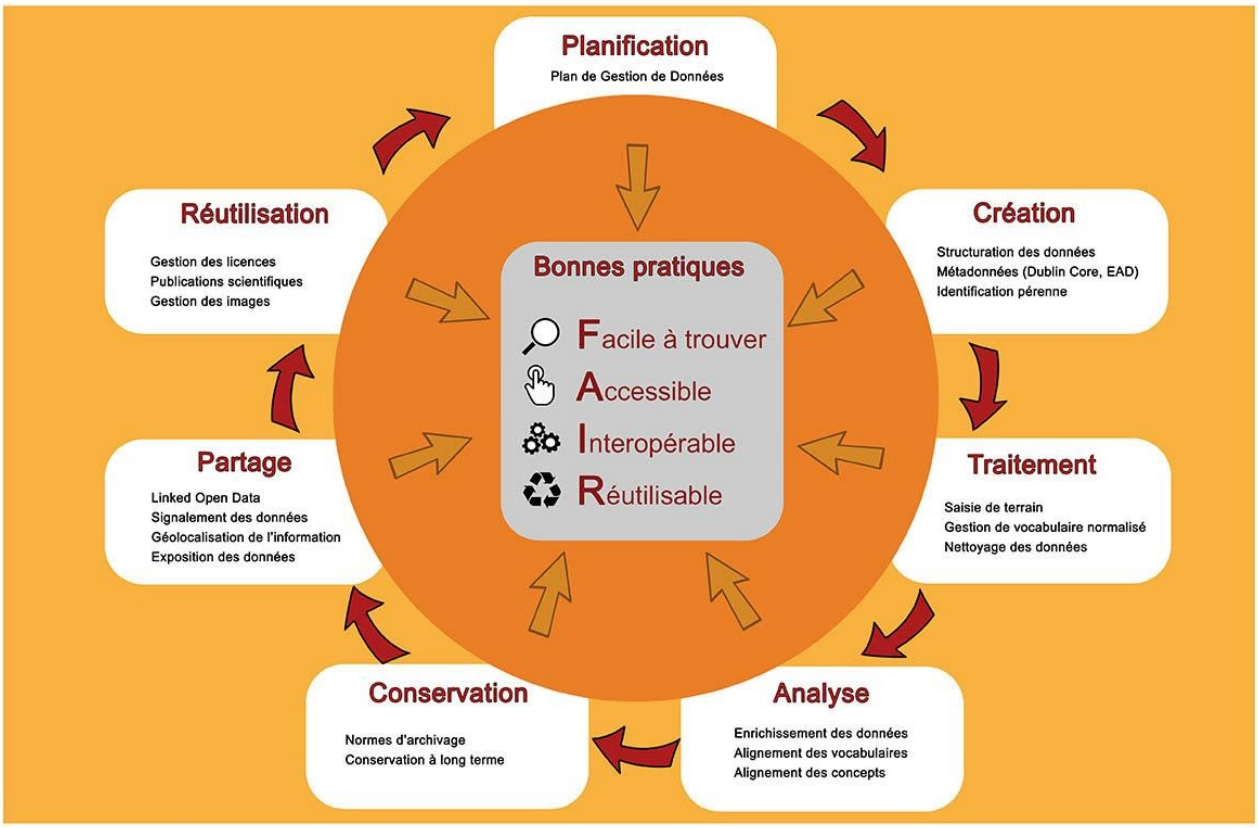


Fig.4 : Le cycle de vie des données du consortium MASA

Ressources :

- Glossaire de la science ouverte proposé par l'URFIST : http://gis-reseau-urfist.fr/flso-glossaire-science-ouverte/

III. Planification

La planification de la gestion des données est une étape indispensable pour identifier les besoins, gagner en efficacité et améliorer la qualité de la recherche. Le plan de gestion des données est un outil essentiel pour organiser les différentes étapes du cycle des données, notamment pour l'intégrité, la préservation et le partage des données de la recherche.

1. Plan de gestion de données

L'établissement du plan de gestion de données (PGD), ou *Data Management Plan*(DMP), est l'occasion d'une réflexion approfondie sur les données, leur structuration mais aussi sur les ressources à mobiliser pour répondre à la problématique de la recherche. La définition de ces besoins permet d'établir les compétences et le budget qui sont nécessaires pour satisfaire chaque étape du projet de recherche.

Dans le cadre de sa politique science ouverte, en lien avec le Plan national pour la science ouverte, l'Agence nationale de la recherche (ANR) demande l'élaboration d'un PGD pour les projets qu'elle finance depuis 2019. Elle participe ainsi à l'alignement européen et international en faveur de la structuration et de l'ouverture des données de la recherche.

2. DMP OPIDoR

La plateforme OPIDoR, hébergée et gérée par l'INIST (CNRS), est un outil d'aide à la création en ligne de plans de

gestion de données (*Data Management Plan* ou DMP) mis à disposition de l'Enseignement Supérieur et de la Recherche. Basé sur le code *open source DMPRoadmap*, il a été adapté aux besoins de la communauté scientifique française.

Le consortium MASA a créé un modèle de plan de gestion de données (PGD) pour l'archéologie sur la plateforme OPIDoR. Fondé sur une version initiale réalisée à l'Inrap en 2018, le PGD MASA a été retravaillé en prenant en compte le modèle proposé par l'ANR et intègre dans ses recommandations les principaux standards et bonnes pratiques applicables à l'archéologie, ainsi que des liens vers une documentation de référence. Le PGD MASA permet de décrire les données de manière globale (onglet Généralités sur les données) puis, plus finement, par jeu de données ou lot de fichiers. Ces ensembles de données peuvent être constitués en référence à une typologie documentaire (photographies, transcriptions, fiches descriptives d'objets, etc.) ou selon des ensembles intellectuellement cohérents (céramologie, anthropologie, etc.). Le modèle organise les informations générales sur le projet ou le contexte de production des données, sur la gestion des données (responsabilités, moyens et procédures qualité mis en œuvre) et sur les politiques d'archivage et de dissémination envisagées à l'issue du projet. Il est disponible en français et en anglais.

Ressources :

- plateforme OPIDoR de l'INIST : https://opidor.fr/planifier/
- PGD MASA sur la plateforme OPIDoR : https://dmp.opidor.fr/public_templates?search=MASA
- Présentation par Marie-Claude Quidoz (CEFE/CNRS) : Plan de gestion des données (PDF)
- Guide OPIDoR par Marie-Christine Jacquemot-Perbal et Laurent Rassinoux : Data management plan ? Plan de gestion de données ? DMP OPIDoR vous guide ! (PDF)
- *Data Management Plan* réalisé los du programme européen PARTHENOS : https://www.parthenos-project.eu/portal/dmp
- Modèle de PGD par l'ANR : https://anr.fr/fileadmin/documents/2019/ANR-modele-PGD.pdf

IV. Création

Avant de collecter les données, il est nécessaire de passer par une étape de modélisation lors de laquelle sont identifiés les principaux éléments qui constitueront le jeu de données et la façon dont ils seront liés les uns aux autres.

1. Structuration des données

La structuration des données ne doit pas être dépendante du choix d'un format ou d'un système. La modélisation doit s'appuyer sur la problématique de recherche et sur les données qui seront mobilisées pour y répondre. Une fois cette modélisation effectuée, peut être posée la question du système et du format les mieux adaptés pour gérer les données .

Le choix du format de stockage des données n'est pas anodin. On oppose, à tort, assez facilement les deux principaux systèmes que sont les SGBD-R (Systèmes de Gestion de Bases de Données Relationnelles) et le XML (*eXtensible Markup Language*), normalisé depuis 1998, bien qu'il existe des SGBD exploitant le XML. Chaque solution présente des avantages selon les cas de figures.

Les principes FAIR recommandent plutôt l'usage du XML, plus ouvert que les SGBD-R. Toutefois, la gestion de corpus volumineux et complexe impose souvent l'usage d'un SGBD-R. Cela n'entrave pas la réutilisabilité des données dans la mesure où il est aisé de proposer un export dynamique des données dans un format XML standard.

Ressources :

- MIAGE (Université de Nantes) : Les SGBD-XML open-sources (site web)
- Comparatif par A. Boukottaya : XML et DBs (PDF)

A. eXtensible Markup Language (XML)

Le langage XML structure l'information à l'aide de balises imbriquées (représentant les entités, voire les caractéristiques) et d'attributs pour préciser l'information et lier certains éléments entre eux. Le HTML par exemple est une exploitation du XML pour le Web. Cette forme d'écriture en arborescence a le mérite d'être lisible aisément par un humain sans nécessiter une application plus complexe qu'un éditeur de texte basique. On privilégie par exemple ce format dès lors que l'information à identifier (on parle de "balisage") se retrouve délayée au sein d'un texte. Ainsi le format XML-TEI permet d'intégrer un balisage au sein d'un texte permettant un enrichissement exploitable ensuite par des applications. Les SGBD-R sont moins adaptés pour ce type d'exploitation de textes.

Ressources :

- Centre de Recherche en Informatique de Lens (Université d'Artois) : Cours XML et XSL (PDF)
- Les outils du Pôle Document Numérique de la MRSH de Caen : Outils XML (EAD et TEI) (site web)

B. Système de Gestion de Bases de Données Relationnelles

Le SGBD-R structure l'information sous forme de tables (représentant les entités) contenant des champs (représentant les caractéristiques) contenant eux-mêmes des valeurs. La principale contrainte des SGBD-R réside dans la nécessité d'utiliser une application pour gérer l'information. On privilégiera les systèmes exploitant le langage normalisé SQL (*Structured Query Language*) qui permet une meilleure portabilité des données vers d'autres systèmes. Il existe des SGBD-R sous forme d'applications de bureau tels que FileMaker ou Microsoft Access mais qui ont le défaut d'être propriétaires et sous licences payantes. De plus, les différentes versions ne sont pas toujours interopérables. Les SGBD-R libres et gratuits, tels que MariaDB (branche libre de MySQL) ou PostgreSQL, permettent de gérer des bases de données en ligne mais ont le défaut d'exiger une installation et un paramétrage qui nécessitent quelques compétences en informatique. En outre, les SGBD-R ont l'avantage de pouvoir gérer rapidement de très gros volumes de données et surtout, dans le cas de MariaDB et PostgreSQL, d'offrir des solutions de publication de ces données. Par la circulaire Ayrault du 19 septembre 2012, le Gouvernement encourage l'utilisation de PostgreSQL. Toutefois, MariaDB est beaucoup plus simple d'utilisation (il ne nécessite pas de connaissances en administration système) et reste le SGBD-R regroupant la plus importante communauté.

Ressources :

- Openclassrooms : Administrez vos bases de données avec MySQL (site web).
- PostgreSQL : Premiers pas avec PostgreSQL (PDF)
- Developpez.com : Cours et tutoriels pour apprendre les SGBD et SQL (site web)

2. Métadonnées (Dublin Core, DCAT, EAD)

Le terme de métadonnées désigne la description des jeux de données partagées, il s'agit en quelque sorte de données sur les données. Trop souvent mal renseignées par les archéologues, quand elles le sont, les métadonnées méritent qu'on prenne le temps de compléter au mieux chaque champ, même si cela peut s'avérer fastidieux. Notamment les noms, la langue du jeu de données, la date de création, les auteur.e.s et leur(s) structure(s) de rattachement, les époques et zones géographiques concernées etc… Les mots clés sont évidemment essentiels ici.

Ces métadonnées permettent à d'autres archéologues de repérer les jeux de données et de choisir ou pas d'utiliser les informations partagées. Les métadonnées sont également moissonnées en premier par des outils de requête comme Isidore-science par exemple.

Ces métadonnées doivent comporter autant d'informations que possible pour faciliter la réutilisation des données : indiquer dans quel cadre un corpus a été mis en place, pour répondre à quelles questions, avec quels moyens, quels outils, toutes ces informations permettront d'évaluer la qualité des données, leur fiabilité et leur pertinence pour un usage dans un autre cadre. Tout enrichissement des données par des ressources externes (Wikipédia, Geonames, etc.) peut également

être mentionné dans les métadonnées afin de préciser l'origine de toutes les informations.

A. Dublin Core

Le Dublin Core (DC) est un vocabulaire du Web sémantique pour décrire les documents ou jeux de données de manière simple et standardisée. Le Dublin Core fournit un socle commun d'éléments descriptifs suffisamment structuré pour offrir une interopérabilité minimum entre des jeux de données variés. Le Dublin Core intègre deux ensemble :

- *DC element set*: 15 propriétés pour décrire de manière simple et générique un document (description, identification, propriété intellectuelle, etc).
- *DC metadata terms*: nombre conséquent d'éléments de description permettant un raffinement très poussé pour décrire un document.

Archeaological Data Service (ADS) est le principal entrepôt numérique pour les données archéologiques du Royaume-Uni. Fondée en 1996, l'activité principale d'ADS est la préservation numérique à long terme des données qui lui sont confiées, avec une politique de gestion et de conservation active des données afin de garantir l'intégrité, la fiabilité et l'accessibilité des données. Toutes les ressources archivées au sein de l'ADS sont en accès libre et sont diffusées sur Internet afin de faciliter leur réutilisation par le secteur du patrimoine et la communauté au sens large. ADS a mis en place un schéma de métadonnées fondé sur le Dublin Core permettant un moissonnage par le protocole OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), protocole permettant d'échanger des métadonnées sur le Web.

Ressources :

- : https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
- Schéma de métadonnées Dublin Core mis en place par l'ADS (PDF - 2003) : http://ads.ahds.ac.uk/arena/DCmeta.pdf (PDF)

B. Data Catalog Vocabulary

Data Catalog Vocabulary (DCAT) est un vocabulaire RDF conçu pour faciliter l'interopérabilité entre les catalogues de données publiés sur le Web. Les éditeurs utilisent le DCAT pour décrire les ensembles de données dans des catalogues en augmentant leur visibilité et en permettant aux applications de les consulter à partir des métadonnées. DCAT permet une publication décentralisée des catalogues qui facilite la recherche fédérée d'ensembles dispersés de données. Les métadonnées DCAT agrégées peuvent servir de fichier associé à un jeu de données pour en faciliter la conservation numérique. C'est ce standard qui a été choisi pour décrire les jeux de données disponibles dans OpenArchaeo.

Ressources :

- Standard de métadonnées interopérables, DCAT : https://en.wikipedia.org/wiki/Data_Catalog_Vocabulary (site web) et https://www.w3.org/TR/vocab-dcat/ (site web)

C. Encoded Archival Description

La description archivistique encodée (EAD) est un format qui utilise le langage XML et permet de structurer des descriptions de manuscrits ou des documents d'archives. Ce modèle comprend des éléments d'identification et d'information relatifs à l'instrument de recherche et au fonds, des éléments de description des composants et sous-composants, des éléments d'informations complémentaires et des éléments d'indexation.

Ressources :

- Guide des bonnes pratiques de l'EAD en bibliothèque : https://www.ead-bibliotheque.fr/guide/ (site web)
- Description de l'EAD sur le site de la BnF : https://www.bnf.fr/fr/ead-encoded-archival-description (site web)
- Traduction française du dictionnaire des éléments de l'EAD version 2002 : https://francearchives.fr/file/0def64f5a10f3f1ae03fdea59399a3e0755ef157/static_1066.pdf (PDF)
- Faire un répertoire ou un inventaire simple en EAD : manuel d'encodage, version 1.1 rédigé par le groupe AFNOR entre juin 2005 et octobre 2009 : https://www.enssib.fr/bibliotheque-numerique/documents/62240-faire-un-repertoire-ou-un-inventaire-simple-en-ead-description-archivistique-encodee.pdf (PDF)
- Informatisation de la description : la DTD EAD sur le site de France Archives : https://francearchives.fr/article/37830 (site web)
- Guide d'indexation pour le Web par le Groupe de travail pour la description et l'indexation des archives sur le Web du service interministériel des archives de France (juin 2021) : https://francearchives.fr/file/6686af73e52bd3dd7d56cbad92228977cbe576f5/GuideIndexation_Web_v202108.pdf (PDF)

D. Norme INSPIRE

En Europe, les données géographiques sont normalisées par la directive INSPIRE, élaborée par la Direction générale de l'environnement de la Commission européenne, qui vise à établir une infrastructure de données géographiques pour assurer l'interopérabilité entre bases de données et faciliter la diffusion, la disponibilité, l'utilisation et la réutilisation de l'information géographique en Europe. Un volet de cette norme est dédié aux métadonnées, notamment celles dédiées à la dimension géographique (système de projection, emprise, échelle…).

De trop nombreux fichiers de relevés topographiques sont inutilisables car les points de référence ou les systèmes de coordonnées n'ont pas été précisés. Le système de projection doit toujours être explicité (Lambert II étendu, WS84, Gauss-Krüger II, etc…), en précisant la version des logiciels utilisés (par exemple Filemaker Pro 7.0.v3 sous Mac, ARCGIS 10.7 sous PC, QGIS 3.10 A Coruña LTR, Adobe Illustrator CS2...).

Ressources :

- Francis Merien et Marc Leobet (2011) : La Directive INSPIRE pour les néophytes (PDF)
- CNIG : Détail de la Directive INSPIRE (site web)
- MSH de Dijon : Catalogue des données géographiques en SHS - CarGOS (site web)

3. Encodage et polices de caractères

Afin d'éviter les difficultés lors d'une réouverture de fichiers, il faut être attentif à l'encodage. C'est ce qui correspond à la manière dont le terminal va interpréter le jeu de caractères utilisé pour l'affichage. Un mauvais encodage peut rendre illisible tous les caractères spéciaux spécifiques à une langue ou à un système. Idéalement, il faut spécifier l'encodage utilisé (dans la plupart des logiciels, cela est paramétrable) dans les métadonnées.

Il faut également prêter une attention particulière aux polices utilisées pour certains fichiers (par exemple Adobe Illustrator ou Photoshop) si celles-ci utilisent des caractères spéciaux qui leur sont propres. *A minima*, il est conseillé d'éviter les polices rares ou protégées (par exemple Helvetica). En cas d'utilisation de polices spéciales pour les alphabets antiques (grec, copte…), il est impératif de créditer précisément la version et le système employé (PC/Mac), si possible en déposant aussi une copie de la police.

Ces conseils de base devraient permettre d'aborder sereinement les vraies questions de sauvegarde et de dépôt : dépôt local ou dépôt national ? Dépôt simple ou dépôts multiples ? Dépôt accessible ou archive longue durée ?

4. Sauvegarde et stockage des données

Le transfert des archives de l'archéologie depuis le papier vers des formats numériques demande beaucoup de rigueur. Lorsque les données sont sauvegardées sous forme numérique, le même soin doit être apporté que celui que met tout archéologue à protéger précieusement les carnets de fouilles, les minutes et esquisses de terrain ou encore les photos.

Il convient d'établir un plan de sauvegarde lié au plan de gestion de données qui précise les différents états dans lequel un fichier doit être sauvegardé (par exemple : scan brut, traitement colorimétrique, version allégée pour le Web), ceci afin d'éviter de stocker de multiples versions inutiles. La sauvegarde peut être doublée et il faut privilégier des types de supports différents et des localisations différentes également (il n'y a aucun intérêt à avoir plusieurs sauvegardes localisées dans un même lieu, au risque de tout perdre en cas de sinistre) : serveur institutionnel, support externe (disque dur clé USB), serveur dédié au stockage (Huma-Num Box par exemple) ou un dépôt en ligne (SEAFILE pour l'Université de Strasbourg, MyCORE pour le CNRS, Zephiryn pour le Ministère de la Culture, etc.). Attention néanmoins à gérer correctement ces sauvegardes multiples en s'assurant de la mise à jour de chacune (versionnement). Chaque type de support a une durée de vie limitée. Il faut donc contrôler régulièrement l'état des sauvegardes et migrer les supports anciens vers des supports plus récents quand cela devient nécessaire.

La question du lieu de stockage est également importante. On évitera les stockages sur des serveurs à l'étranger pour préférer le national. On évitera le stockage chez des prestataires privés pour privilégier les infrastructures institutionnelles (bien que certaines de nos institutions font appel à des prestataires privés...).

5. Stockage de données à caractère personnel

En cas de stockage de données à caractère personnel, que ces informations soient publiées ou non, il importe de bien respecter les réglementations en vigueur (Règlement Général sur la Protection des Données).

Ressources :

- InSHS : La protection des données à caractère personnel dans le contexte de la science ouverte – Guide pour la Recherche (PDF)
- Explication de la RGPD par la CNIL : https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on (site web)
- Règlement Général sur la Protection des Données (CNIL) : https://www.cnil.fr/fr/reglement-europeen-protection-donnees (site web)

V. Traitement

Le traitement des données a pour objectif d'assurer l'intégrité des données saisies et à les préparer efficacement pour leur analyse. Plus le jeu de données est structuré rigoureusement avec des contraintes de saisie et des listes de choix fermées, moins le nettoyage est nécessaire. À l'inverse, il est plus long de nettoyer un jeu de données constitué de nombreux champs descriptifs ouverts sans contraintes de saisie.

Pour tous les traitements qui concernent des jeux de données 3D, il est conseillé de se référer aux travaux du consortium 3D-SHS et notamment au Livre Blanc du consortium 3D-SHS.

1. Nettoyage des données

Cette opération consiste à repérer et homogénéiser dans un jeu de données des termes identiques dont l'orthographe pourrait varier au sein de la base pour différentes raisons : fautes de frappe, inversion de mots, conjugaison d'un verbe, terme au pluriel, présence d'espaces superflus, etc. Cette étape est indispensable pour engager des analyses mais aussi pour l'alignement des données avec des référentiels (vocabulaires normalisés) nécessaires pour l'interopérabilité, le partage et la publication dans le Web des données. Le nettoyage consiste à vérifier, et le cas échéant à corriger, que les données se présentent dans un format uniforme et normalisé.

OpenRefine, logiciel libre et gratuit, permet de réaliser le nettoyage, la préparation et l'enrichissement des données (cf VI.2.A).

Ressources :

- Site officiel d'OpenRefine : https://openrefine.org/ (site web)
- Mathieu Saby (2020) : Tutoriel OpenRefine (site web)

2. Mise en oeuvre d'un vocabulaire normalisé

La création d'un vocabulaire normalisé ou référentiel est une étape importante dans la vie d'un projet scientifique, qui fait appel à des compétences variées : le chercheur apporte sa connaissance du domaine, le documentaliste sa maîtrise des formats des métadonnées et de l'indexation, l'informaticien son expertise dans la mise en oeuvre des outils et des systèmes de gestion et d'alignement des vocabulaires normalisés. Ces trois compétences se révèlent essentielles à la réussite de l'opération.

Il n'existe que très peu de logiciels open source pour la gestion de vocabulaires respectant la norme ISO 25964. Opentheso est l'un de ses logiciels. Lancé en 2005 par la plateforme WST (Web Sémantique et Thesauri) de la Maison de l'Orient et de la Méditerranée (Lyon), son développement est soutenu par le réseau FRANTIQ, le consortium MASA et l'IR* Huma-Num. À la suite de la sortie de la nouvelle norme de thésaurus en 2011, une refonte complète du logiciel a été engagée avec les contraintes suivantes :

- respect de la norme ISO 25964
- production d'une version fullweb
- soin de l'interface graphique avec un accent particulier pour la convivialité et la simplicité
- diffusion du logiciel en open source

Opentheso représente actuellement une plateforme complète de gestion collaborative et de normalisation de thésaurus, il est composé de trois grandes briques (Fig. 5) :

- Une interface web qui permet de centraliser, gérer, normaliser, aligner des centaines de thésaurus sur une seule instance, le travail se fait en mode collaboratif.
- L'attribution d'identifiants pérennes de type Ark et Handle, l'édition des thésaurus aux formats normalisés pour le Web de données et pour l'échange avec d'autres partenaires (RDF, SKOS, Json, JsonLd, Turtle, PDF, CSV)
- Une API complète de type REST qui permet d'exploiter les vocabulaires à distance en utilisant des applications métiers adaptées à chaque usage, par exemple l'écosystème numérique du chantier de reconstruction de Notre-Dame de Paris, le CMS OmekaS, le système intégré de gestion de bibliothèque Koha, la chaîne d'édition numérique Métope… (fig. 6).

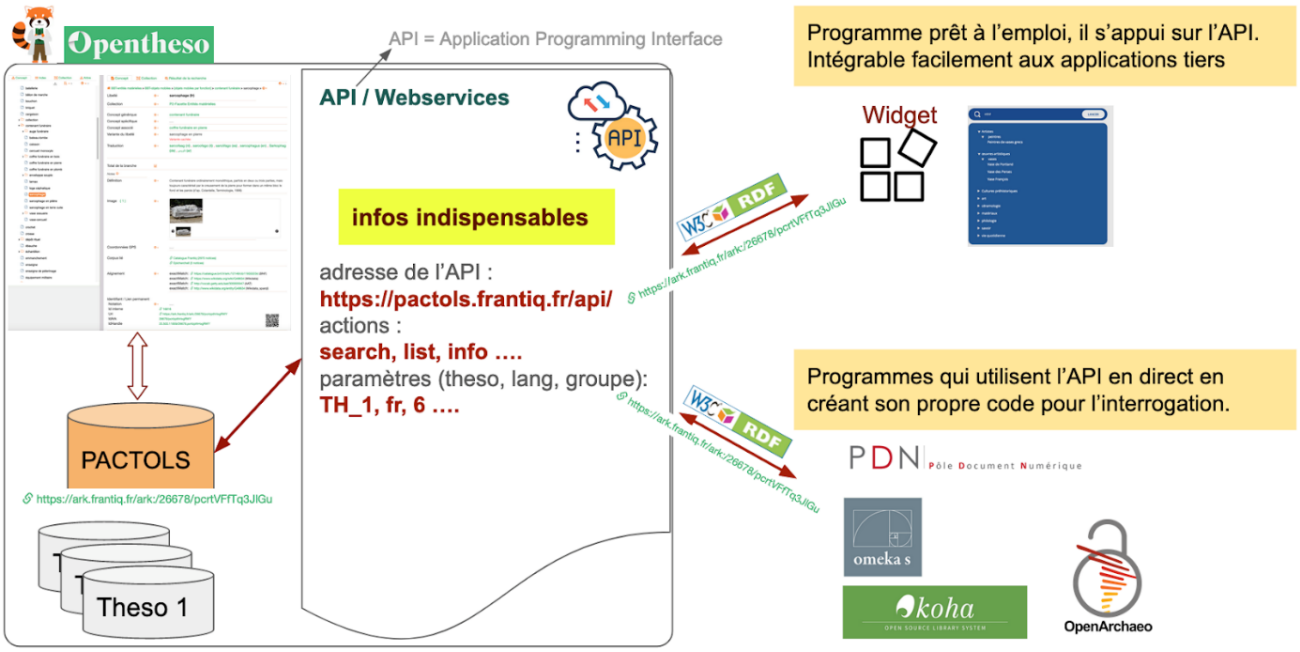


Fig. 5 : Les 3 briques composant Opentheso : interface de gestion et de consultation ; normalisation et identification pérenne ; l'API d'interconnexion de type REST avec échange des données en RDF.

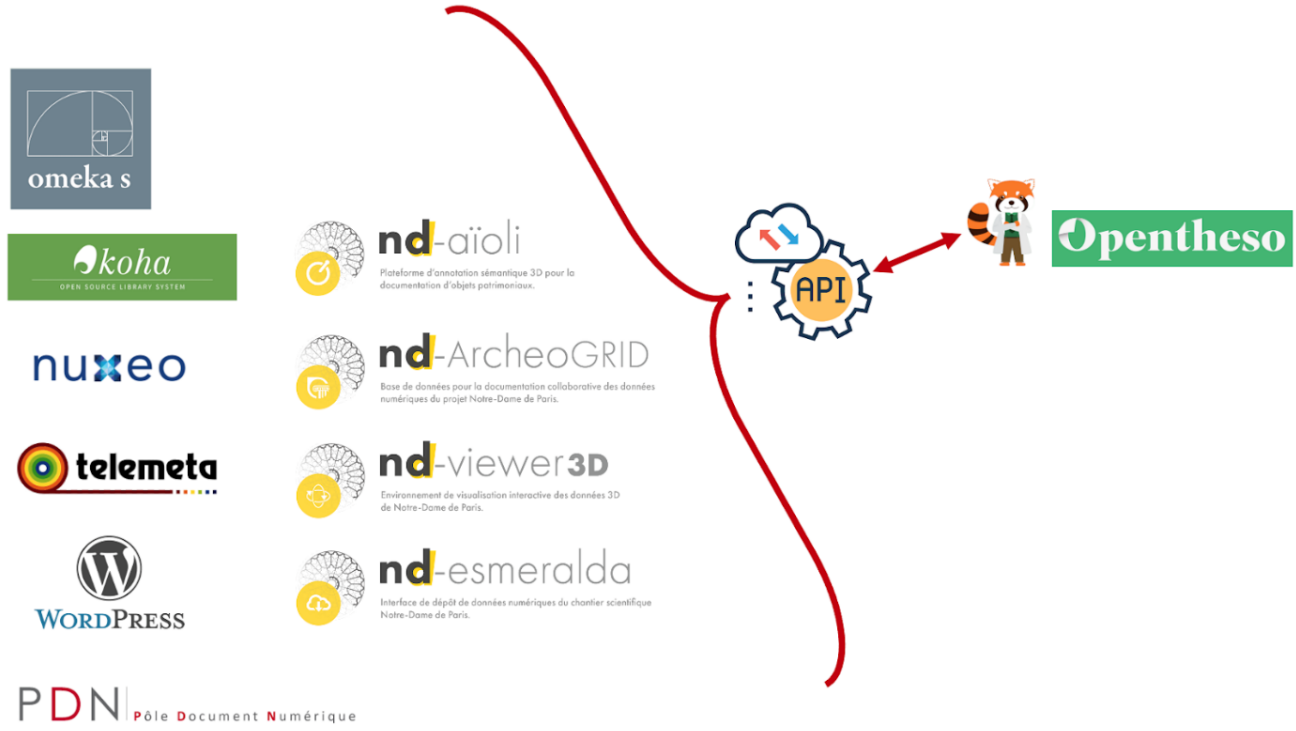


Fig. 6 : Quelques exemples d'applications et projets numériques exploitant Opentheso

Ressources :

- Blog présentant OpenTheso : https://opentheso.hypotheses.org/ (site web)
- code source : https://github.com/miledrousset/Opentheso2 (site web)

VI. Analyse

L'analyse archéologique des données relève de chaque problématique de recherche et est de la responsabilité des chercheurs, indépendamment de toute question numérique. Néanmoins, l'enrichissement des données descriptives en s'appuyant sur des référentiels connus et validés par les communautés est un préalable au bon déroulement des analyses. Par exemple, des champs concernant les personnes dans les métadonnées (dc:creator, dc:subject) devraient utiliser les labels d'un des référentiels IdRef, VIAF, ORCID, voire AuréHAL afin que les noms respectent une forme valide et résolvent d'éventuels cas de synonymie. Ces référentiels sont liés et proposent des enrichissements réciproques (cf. VI.1. B). De même, pour les champs destinés à décrire le contenu de la ressource (dc:subject ou équivalents), il est préférable d'utiliser des listes de termes validés et partagés par la communauté. À ce titre, le thésaurus PACTOLS, développé par Frantiq et soutenu par MASA, couvre les besoins de mots-clés du vocabulaire archéologique.

L'objectif de publication dans le Web de données implique une analyse sémantique approfondie qui consiste à rapprocher et à aligner autant que possible les vocabulaires utilisés avec des référentiels normalisés, puis les concepts avec ceux d'ontologies partagées par la communauté. On assure ainsi l'interopérabilité des données à deux niveaux :

- en associant autant que possible les données ou les métadonnées avec des référentiels partagés par la communauté scientifique ;
- en proposant un alignement sémantique des données avec une ontologie, elle aussi partagée par la communauté, en l'occurrence pour l'archéologie : le CIDOC CRM (cf. VI.4).

1. Analyse des données et indexation

La description d'une ressource gagnera à s'appuyer sur des listes de termes choisis et validés *a minima* par les chercheurs qui utilisent la base de données. Est établie alors une sélection de termes à aligner sur un référentiel terminologique plus largement répandu. Par exemple, la base de données du projet EpiCherchell a défini des index avec des listes fermées de mots-clés à l'intitulé personnalisé, mais chaque terme est associé à un concept équivalent du thésaurus PACTOLS par ajout de son identifiant ARK. Dans la perspective d'une publication en ligne et d'un partage de ces données, il est recommandé d'utiliser directement des vocabulaires normalisés reconnus par une communauté élargie. L'interopérabilité implique un choix raisonné du vocabulaire de description des données. Un lexique commun favorise la compréhension, facilite la recherche, surtout si celle-ci doit se faire dans un réservoir qui accueille des ressources/jeux de données d'origine différente. Le vocabulaire constitue alors un atout pour l'analyse.

A. Mots-clés pour décrire une ressource

Le choix des mots-clés pertinents doit répondre à la question : de quoi parle cette ressource ? Ce choix doit également permettre, lors d'une requête ou de la consultation d'un index, de produire des résultats qui excluent des réponses inappropriées, ou au contraire, qui évitent l'absence d'une donnée pertinente dans la ressource.

En archéologie, les mots-clés doivent rendre compte des sujets, personnes, lieux et périodes ou dates caractérisant le jeu de données. La personne qui recherche une information n'a pas les présupposés de celle qui a renseigné la base de données. Il est donc recommandé qu'une personne différente procède régulièrement à des contrôles de saisie et à des sondages dans le jeu de données pour en contrôler la cohérence. Le nombre de mots-clés ajoutés doit être raisonnable si le champ d'indexation autorise la saisie multiple.

Ressources :

- SCEREN. MOTBIS 2010 : guide d'indexation [http://www.cndp.fr/motbis/telechargement/guide_d_indexation.pdf]. CNDP, 2010 (PDF)

B. Référentiels

Les référentiels présentés ci-dessous proposent tous des notices d'autorité ou des vocabulaires de référence respectueux des normes internationales. La plupart offrent de plus des services d'export ou d'interconnexion (APIs et *webservices*) dans le but de diffuser largement et de partager leurs données.

LOTERRE est une plateforme de publication en ligne et de partage de vocabulaires normalisés scientifiques, multidisciplinaires et multilingues mis en place par l'INIST. Pour l'archéologie elle propose des ressources en Géographie, Art et Archéologie, Préhistoire et Protohistoire, Histoire et sciences des religions, Philosophie. Ces vocabulaires sont figés depuis l'arrêt des bases de données bibliographiques Francis (pour l'indexation de la littérature européenne en lettres et sciences humaines et sociales) en 2015.

a. Référentiels thématiques

Référentiels thématiques

Art & Architecture Thesaurus (AAT) est le thesaurus développé par la fondation Getty consacré à l'art, l'architecture et le patrimoine culturel. Utilisé comme standard documentaire, il fait autorité et est largement utilisé par la communauté internationale. Ouvert et lié dans le *Linked Open Data*, il est diffusé sous licence libre (ODC-By) 1.0. L'infrastructure européenne ARIADNE utilise l'AAT comme thésaurus pour les champs thématiques en archéologie.

Ressources :

- Tutoriel de Lizzie Scholtus (MASA) : Alignement de vocabulaire avec Getty's AAT

PACTOLS est le thésaurus créé et développé par le réseau Frantiq. Sa structure normalisée organise et relie par des relations sémantiques riches env. 60.000 concepts ou mots-clés au sein de plusieurs collections thématiques qui couvrent tous les domaines de l'archéologie et des sciences de l'Antiquité. Il est exploité dans des contextes documentaires (catalogues de bibliothèques, inventaires archivistiques), éditoriaux, et plus récemment pour l'enrichissement de données de recherche ou de valorisation de ressources patrimoniales. Chaque concept est traduit en au moins 6 langues et porte un identifiant pérenne ARK qui permet de le lier à tout vocabulaire d'indexation propre à un système de base de données ou à un référentiel en ligne. Il est aligné principalement sur Wikidata, IdRef, GeoNames, AAT. Il est diffusé sous licence libre ODbL.

Ressources :

- Tutoriel de Lizzie Scholtus (MASA) : Alignement de vocabulaires avec les PACTOLS

Wikidata est la base de référence des projets de la *Wikimedia Foundation*. Ses items servent de pivots de référence aux ressources de Wikipedia ou de Wikicommons par exemple. Elle est alignée avec de nombreux référentiels dans tous les domaines de la connaissance. Sa langue de référence est l'anglais.

Quel vocabulaire choisir ? MASA soutenant les PACTOLS, le choix de ce thésaurus est évident pour tout corpus archéologique national. Toutefois, participant à l'infrastructure européenne ARIADNEplus, MASA préconise de procéder également à l'alignement des données avec l'AAT pour assurer l'interopérabilité des données également au niveau international. En outre, multiplier les alignements à divers référentiels, bien que cela représente un temps non négligeable, ne peut pas nuire et il peut aussi être pertinent d'aligner également le vocabulaire avec Wikidata. Le programme européen DARIAH met actuellement en place un thésaurus international nommé Backbone Thesaurus (BBT) qui permet d'interconnecter l'ensemble des thésaurus en ligne pour les sciences humaines et sociales, dont l'AAT et les PACTOLS.

b. Référentiels géographiques

Référentiels géographiques

GeoNames est un référentiel géographique permettant d'identifier les lieux en tant que concepts. Il attribue un identifiant unique et pérenne à un lieu à quelque échelle qu'il soit (monument, hameau, commune, agglomération, département, région, pays). Il s'agit d'un thésaurus multilingue qui permet de gérer l'arborescence des différents concepts. Plusieurs services Web sont mis à disposition pour exploiter cette ressource : récupération d'un concept à partir du code postal ou de coordonnées latitude/longitude ou à l'inverse récupération des coordonnées géographiques à partir de l'identifiant, par exemple.

Il est possible de compléter le référentiel pour les entités de bases (monument, lieu-dit, commune) simplement en se créant un compte. Pour des entités plus complexes (une région historique qui n'a plus de réalité administrative aujourd'hui par exemple), il faut contacter les responsables pour soumettre la création d'un assemblage d'entités existantes.

Ressources :

- Geonames : http://www.geonames.org/ (site web)
- Abdelmajid Khayari, Gilles Banzet : Retour d'expérience sur l'alignement d'un thésaurus sur GeoNames (2019) (PDF)
- Tutoriel de Lizzie Scholtus (MASA) : Alignement de vocabulaire de localisation avec GeoNames

c. Référentiels chronologiques

Référentiels chronologiques

PeriodO, choisi par l'infrastructure européenne ARIADNEplus, est un répertoire de définition de périodes chronologiques (histoire, archéologie, histoire de l'art) différentes selon les aires culturelles. Il facilite l'établissement de lien entre les chronologies dont il permet d'identifier les chevauchements. L'Inrap, par exemple, a versé la chronologie de référence utilisée par l'établissement.

Ressources :

- PeriodO : http://perio.do/ et chronologie utilisée par l'Inrap (site web)
- L'ontologie du temps de la W3C (2020) : https://www.w3.org/TR/owl-time/ (site web)
- Chronontology de la German Research Foundation (2015) : https://chronontology.dainst.org/ (site web)

d. Référentiels auteur/individu/institution

Référentiels auteur/individu/institution

Ces référentiels sont souvent produits par des bibliothèques ou des centres documentaires qui en garantissent le contrôle et une mise en relation par des alignements réciproques. Leur objectif initial est l'attribution sans ambiguïté d'une œuvre à son ou ses auteurs. Le choix du référentiel dépend des entités à traiter, de leur typologie et de leur origine française ou étrangère.

Certains de ces référentiels ajoutent aux notices d'autorité les identifiants d'autres référentiels, de façon automatique pour ceux qui sont alignés ou manuellement pour ceux qui ne le sont pas.

AuréHAL est l'Accès Unifié aux Référentiel HAL. Les autorités de l'archive ouverte HAL concernent les auteurs, les organismes académiques et de recherche relatifs aux articles, chapitres d'ouvrages ou rapports déposés. C'est une ressource essentielle pour identifier des auteurs qui ne sont pas repérés dans les catalogues de bibliothèques.

Le VIAF (*Virtual International Authority File* / Fichier d'autorité international virtuel) est un projet commun à plusieurs bibliothèques nationales mis en œuvre et hébergé par OCLC. Leur but est la mise en commun de leurs fichiers d'autorités permettant des économies de moyens et une valorisation par l'alignement des fiches d'autorité. L'opérateur français du VIAF est la BnF. Ses autorités concernent les personnes, institutions, noms géographiques, titres d'œuvres, notices d'autorités Rameau et Dewey (classifications bibliothéconomiques). Il agrège les fichiers d'autorité de plusieurs fournisseurs français dont l'Abes qui produit IdRef.Le VIAF signale principalement des auteurs d'ouvrages, de travaux universitaires, de rapports habituellement signalés dans les catalogues de bibliothèques, rarement d'articles ou de chapitres d'ouvrages, les bibliothèques pratiquant rarement le dépouillement de travaux collectifs.

Ressources :

- Tutoriel de Lizzie Scholtus (MASA) : Alignement de vocabulaire avec VIAF

IdRef (Identifiants et Référentiels pour l'enseignement supérieur et la recherche) rassemble les autorités produites par l' Abes, Agence bibliographique de l'enseignement supérieur, et ses partenaires académiques qui produisent ensemble le catalogue du système universitaire de documentation (Sudoc).Comme le VIAF, IdRef référence des auteurs de monographie, mais l'intégration dans le réseau Sudoc de bibliothèques de recherche qui effectuent souvent le dépouillement d'ouvrages collectifs ajoute des auteurs de contributions aux ouvrages scientifiques.

ORCID (*Open Researcher and Contributor ID*) est une organisation sans but lucratif dont l'objectif est la mise en réseau des chercheurs de toutes les disciplines par le biais d'un identifiant pérenne personnel unique et gratuit. Chacun peut demander l'attribution d'un identifiant, il n'y a pas de contrôle sur les données déclarées. Cet ID sert à lier la fiche individuelle à des ressources comme les archives ouvertes. Son succès en fait une référence qui est demandée par les éditeurs et les agences de financement pour identifier le chercheur.Le service étant déclaratif, toute personne, même n'ayant jamais publié, peut demander l'attribution d'un ID mais un auteur décédé ne peut donc pas bénéficier d'identifiant ORCID.

2. Alignement de vocabulaires

L'alignement de vocabulaire consiste à rapprocher les mots-clés ou descripteurs utilisés dans un jeu de données avec un référentiel, de préférence en ligne et partagé par une communauté. Le lien s'effectue à l'aide d'un identifiant pérenne (identifiant unique accessible en ligne via une adresse web), soit au sein même du jeu de données soit dans ses métadonnées.

Il s'agit de sélectionner dans les données sources les termes à aligner en fonction de leur valeur scientifique – c'est-à-dire de l'intérêt qu'il y a à signaler cette information ou à la rechercher – mais aussi selon le modèle de données du catalogue ou de l'entrepôt dans lequel va s'inscrire la base référencée. Ces termes sont ensuite mis en correspondance avec ceux du référentiel à l'aide d'outils d'alignement (Vocabulary Matching Tool, le service de réconciliation d'OpenRefine ou en lui indiquant l'API d'un service de référentiel). Le module d'alignement d'Opentheso est pré paramétré pour aligner les vocabulaires avec les sources suivantes : Wikidata, IdRef, Getty AAT, Geonames, Pactols, Gemet, Agrovoc. Pour que cette mise en correspondance puisse fonctionner automatiquement, il est nécessaire que les termes sélectionnés aient été nettoyés au préalable (cf. V.1). Toutefois, si ces outils permettent d'automatiser une partie du travail d'alignement, il reste nécessaire de désambiguïser le résultat en contrôlant les termes du référentiel obtenus par alignement. Il s'agit de vérifier s'il n'y a pas de confusion entre des homonymes et de contrôler la position des termes dans l'arborescence du thésaurus pour éviter tout contresens. Il est également utile de consulter les synonymes possibles des termes pour lesquels aucune équivalence n'aurait été déterminée. La validation doit être effectuée par le producteur des données, ou la personne qui les maîtrise le mieux.

Lorsque l'alignement entre les données et le référentiel choisi est confirmé, il est recommandé d'ajouter l'identifiant pérenne dans la base ou dans ses métadonnées.

A. OpenRefine

Outre le nettoyage des données (cf. V.1), OpenRefine peut servir de passerelle pour convertir les données dans différents formats. Il propose également des services dit de "réconciliation" pour enrichir les données en exploitant les services web de référentiels tels que Wikipédia, PACTOLS, GeoNames, par exemple.

Ressources :

- OpenRefine : https://openrefine.org/ (site web)
- Tutoriel OpenRefine par Mathieu Saby : https://msaby.gitlab.io/tutoriel-openrefine/ (site web)
- https://programminghistorian.org/fr/lecons/nettoyer-ses-donnees-avec-openrefine : leçon du *Programming Historian* pour le nettoyage de données avec OpenRefine (site web)

B. Vocabulary Matching Tool

L'infrastructure européenne ARIADNEplus propose, parmi les outils qu'elle met à disposition de la communauté, l'application en ligne *Vocabulary Matching Tool* qui permet d'aligner des vocabulaires avec le AAT (cf. VI.1.B.a).

Ressources :

- : https://heritagedata.org/vmt2/vmt-app.html (site web)
- Tutoriel sur la plateforme ARIADNEplus : Vocabulary Matching Tool (site web)

C. Opentheso

Opentheso (cf. V.2) dispose d'un module d'alignement intégré qui permet de réaliser l'alignement depuis la notice du concept. Plusieurs référentiels sont déjà intégrés qu'il suffit d'appeler pour procéder à un alignement automatique. Si le référentiel visé n'a pas été enregistré, l'alignement peut être effectué manuellement. Attention, l'alignement automatique s'appuie sur le libellé préférentiel du concept dans la langue de référence du thésaurus de départ. Le repérage d'un item correspondant sera donc plus délicat à réaliser si la langue de référence du référentiel visé est différente.

Ressources :

- https://opentheso.hypotheses.org/tag/alignement (site web)
- Alignements automatiques et manuels : https://opentheso.hypotheses.org/52 (site web)

3. Enrichissement des données

L'enrichissement des données consiste à importer, dans un jeu de données source, des informations provenant d'un référentiel en ligne partagé. Il s'agit donc d'une opération qui s'effectue après un alignement de vocabulaire. Tous types d'éléments peuvent ainsi être récupérés pour compléter des données.

Le logiciel OpenRefine permet cet enrichissement grâce à son service de réconciliation. Il est possible, après avoir aligné les termes par réconciliation, de récupérer les identifiants pérennes du référentiel directement et toutes autres informations en utilisant quelques lignes de commandes.

Certains référentiels proposent aussi des services web (Opentheso par exemple) permettant de moissonner tout ou partie des informations disponibles pour un terme. Il est parfois nécessaire de nettoyer ou retravailler les éléments ainsi récupérés pour les intégrer correctement dans le jeu de données.

Ressources :

- https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine : leçon du *Programming Historian* sur le moissonage et nettoyage de données en ligne (site web)
- https://www.getty.edu/research/tools/vocabularies/obtain/getty_vocabularies_openrefine_tutorial.pdf : tutoriel pour faire la réconciliation avec le Getty AAT dans OpenRefine (PDF)
- https://docs.openrefine.org/manual/reconciling : manuel de OpenRefine pour utiliser le service de réconciliation (site web)

4. Alignement conceptuel

Les concepts mobilisés dans les nombreuses bases de données archéologiques sont bien souvent similaires de l'une à l'autre, mais souvent avec des dénominations différentes et avec des relations entre les concepts qui peuvent varier. Cela rend impossible une interrogation commune des différentes bases, d'autant plus qu'elles fonctionnent dans des formats différents (mySQL, postgreSQL, Filemaker, MS Access, 4D). La solution d'interopérabilité retenue et recommandée par le consortium MASA, à l'instar de nombreux projets européens (ARIADNEplus, PARTHENOS, 4CH…), consiste à passer par un langage commun pour apparier les concepts de chaque base de données avec ceux d'une ontologie.

L'ontologie sert alors de surcouche commune aux différents jeux de données auxquels elle fournit une grammaire unique pour interroger conjointement sans avoir à modifier ni la structure ni les vocabulaires des bases initiales. L'ontologie retenue pour l'archéologie est celle du CIDOC CRM, dédiée au domaine du patrimoine culturel.

A. L'ontologie du CIDOC CRM

a. Qu'est-ce que le CIDOC CRM ?

Qu'est-ce que le CIDOC CRM ?

Le CIDOC CRM (Comité international pour la documentation [réseau de l'*International council of museums*, ICOM]-*Conceptual Reference Model*) est une ontologie de domaine qui permet de construire des modèles de données compatibles avec le Web sémantique. Le CIDOC CRM propose une représentation conceptuelle implicite et explicite du domaine du patrimoine culturel, dans l'objectif d'en partager et promouvoir les données selon un standard sémantique commun. Il est composé de classes et de propriétés inférées destinées à représenter tous les objets, toutes les informations et tous les concepts utilisés dans le domaine du patrimoine culturel. L'ontologie tend vers une logique sémantique, propre à la recherche et aux inventaires du patrimoine culturel, destinée aux machines afin de favoriser l'interopérabilité et l'interrogation des différents jeux de données.

L'ontologie est maintenue et enrichie par un *Special Interest Group* (CRM-SIG). Des extensions complètent l'ontologie initiale pour couvrir par une meilleure description certains sous-domaines du patrimoine culturel (archéologie, bâti, bibliothèque, etc…).

Ressources :

- Le site du CIDOC CRM : http://www.CIDOC CRM.org/ (site web)

b. Un modèle standard commun

Un modèle standard commun

D'abord en décembre 2006, puis en 2014, le CIDOC CRM a été certifié ISO (ISO-21127:2014), devenant ainsi la norme internationale pour la documentation et l'échange d'informations concernant le patrimoine culturel.

Les modèles d'appariement de chaque jeu de données peuvent être hétérogènes, malgré l'utilisation d'une ontologie commune. Leur interrogation conjointe reste une difficulté. Pour la contourner, le consortium MASA propose un modèle générique d'appariement des jeux de données archéologiques avec le CIDOC CRM. Il décrit les inventaires de sites, les archives de fouilles ou les collections de mobilier archéologique avec une granularité assez large pour croiser des jeux de données variés et suffisamment fine pour les distinguer. L'objectif de cette approche est de mettre en œuvre le modèle pour publier des corpus archéologiques dans le Web des données et de l'implémenter derrière une interface d'interrogation simplifiée. La plateforme OpenArchaeo (cf. VIII.2.B) a été développée selon les principes du Web sémantique : les données sont stockée dans un triplestore RDF (cf. VIII.2.A) structuré selon l'ontologie CIDOC CRM en mobilisant des référentiels normalisés (PACTOLS , AAT, GeoNames, VIAF) et peuvent être interrogées par un langage propre au Web sémantique (SPARQL). Elle offre une interface d'interrogations simplifiées à partir de concepts archéologiques partagés.

Ressources :

- Modèle de données générique en CIDOC CRM pour OpenArchaeo : http://openarchaeo.huma-num.fr/explorateur/home (site web)

B. Outils d'appariement de concepts

Pour associer les jeux de données d'un format classique (XML, SQL) à OpenArchaeo ou tout autre triplestore RDF, il est nécessaire d'apparier les concepts utilisés avec les classes (ou "entités") du modèle générique. En fonction du jeu de données, il peut être nécessaire de procéder au préalable à l'alignement des vocabulaires avec des référentiels (cf. ~~VI.~~2),

revoir le format de certaines données ou réorganiser l'information (cf. V.1). Pour l'appariement des données, si celles-ci sont dans un SGBD-R, on utilisera Protégé-Ontop ; si elles sont en format XML, on utilisera *Mapping Memory Manager*.

a. Protégé-Ontop

Dans le cas d'un SGBD-R, les données qui correspondent à une entité du modèle générique sont contenues dans les champs du SGBD-R. Dans ce cas, l'opération d'appariement (mapping en anglais) entre le contenu d'un champ de base de données et un concept (ou classe, ou entité) de l'ontologie (CIDOC CRM) s'effectue avec l'application Ontop, la connexion avec la donnée se faisant avec le langage d'interrogation propre au SGBD-R (SQL).

Protégé-Ontop est une combinaison du logiciel Protégé et de son plugin Ontop. Protégé-Ontop permet de préparer des requêtes SQL pour récupérer dynamiquement des données venant d'une base de donnée, d'éditer un appariement de données avec une interface utilisateur au sein du logiciel et de lancer la transformation des données pour obtenir un fichier compatible avec le Web sémantique (RDF-XML, Turtle).

Ontop est un projet qui dépasse le seul plugin pour Protégé et existe dans d'autres formats. Néanmoins, seule la solution avec Protégé offre une interface utilisateur pour réaliser les appariements.

L'utilisation de Protégé-Ontop nécessite l'apprentissage du logiciel Protégé, qui sert à éditer des ontologies. Pour réaliser un appariement, il faut d'abord utiliser cette fonction de Protégé pour préparer l'ontologie utilisée par le modèle de données afin que toutes les classes et propriétés utilisées soient accessibles au plugin Ontop. Toutes les classes et propriétés utilisées sont enregistrées dans un fichier OWL réutilisable pour d'autres appariements.

Une fois l'ontologie prête à l'utilisation, le plugin Ontop propose de nouvelles fonctions à l'utilisateur, dont la connexion avec une base de données SQL (MySQL ou PostGreSQL). L'opération consiste à installer un driver de connexion vers le SGBD utilisé, comprenant l'URL et les identifiants d'accès à la base.

Une fois la connexion établie, l'utilisateur peut définir les correspondances entre la base de données et l'ontologie et formaliser ainsi les liens entre le jeu de données et l'ontologie. L'ensemble de ces liens sont stockés dans un fichier OBDA (Ontology Based Data Access) réutilisable (par exemple interrogation dynamique du jeu de données à partir de l'ontologie ou mise à jour des données que l'on veut porter sur le Web sémantique).

Ressource :

- Ontop : https://ontop-vkg.org/ (site web)
- Tutoriel en anglais : https://ontop-vkg.org/tutorial/ (site web)
- Tutoriel en français (MASA) : Aligner une base de données avec une ontologie avec Ontop (PDF)

b. Mapping Memory Manager (3M)

Dans le cas d'un jeu de données XML, les données qui correspondent à une entité du modèle générique sont contenues dans les balises du fichier XML. Dans ce cas, l'opération d'appariement (*mapping*en anglais) entre le contenu d'une balise et un concept (ou classe, ou entité) de l'ontologie (CIDOC CRM) s'effectue avec l'application en ligne Mapping Memory Manager (3M), les liens avec la donnée se faisant avec le langage X-Path.

3M est un outil d'appariement en ligne développé par la FORTH. Il permet de créer des appariements entre des fichiers XML et des ontologies, notamment le CIDOC CRM inclus nativement dans 3M. 3M propose une interface utilisateur pour X3ML, un langage réservé à l'alignement de nœuds XML avec une classe ou un type de donnée littérale d'une ontologie, afin de créer des triplets RDF à partir de XML. Pour récupérer les données contenues dans une balise XML ou dans un attribut, on localise son emplacement dans le fichier à l'aide de Xpath puis on l'associe à une classe de l'ontologie CIDOC CRM.

Une fois l'appariement réalisé, la fonction Transformation de 3M permet, à partir du jeu de données XML, de générer le fichier contenant les triplets RDF (au format RDF-XML, Turtle, nTriple).

Ressources :

- OLDMAN Dominic, THEODORIDOU Maria, SAMARITAKIS Georgio, Using Mapping Memory Manager (3M) with CIDOC CRM (PDF)
- Tutoriel de Florian Hivert (MASA), Mapping de données avec 3M (PDF)
- Guide de prise en main (anglais) : http://139.91.183.3/3M/Manuals/Help.pdf (PDF)
- Manuel d'utilisation de 3M (anglais) : http://139.91.183.3/3M/Manuals/en/manual.pdf (PDF)
- Tutoriel pour construire les générateurs d'instances et d'étiquettes (anglais) : http://139.91.183.3/3M/Manuals/en/X3ML_Generators_Manual.pdf (PDF)
- Tutoriel d'initiation à 3M par George Bruseker (2019) : https://masa.hypotheses.org/files/2019/10/X3MLToolkitTutorial.pdf (PDF)

VII. Conservation

La conservation des données associe leur archivage et leur préservation. L'archivage implique de rendre intelligible le jeu de données en exploitant ses métadonnées. La préservation consiste à garantir l'accès aux données dans le temps. Quand la préservation est prévue pour être à long terme, on parle d'archivage pérenne.

Pour la préservation des données 3D, fréquentes en archéologie, il est conseillé de se référer aux bonnes pratiques publiées par le consortium 3D-SHS et notamment au générateur d'archive 3D pérenne aLTAG3D et au Conservatoire Nationale des données 3D.

Ressources :

- Les recommandations du Consortium 3D-SHS : https://hal.archives-ouvertes.fr/hal-01683842/ (site web)

1. Normes d'archivage

Le Code du patrimoine définit les archives comme l'ensemble des documents quels que soient leur date, leur lieu de conservation, leur forme et leur support produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité. Ce texte fait notamment une distinction entre les archives publiques et les archives privées. En effet, les archives publiques comprennent les documents qui procèdent de l'activité de l'Etat, des collectivités territoriales, des établissements et entreprises publics ; les documents qui procèdent de l'activité des organismes de droit privé chargés de la gestion des services publics ou d'une mission de service public ainsi que les minutes et répertoires des officiers publics ou ministériels. Les archives privées sont définies par opposition aux archives publiques comme l'ensemble des documents qui n'entrent pas dans le champ d'application des archives publiques.

La loi n°78-753 du 17 juillet 1978 dite loi CADA définit comme documents administratifs les dossiers, rapports, études, comptes-rendus, procès-verbaux, statistiques, directives, instructions, circulaires, notes et réponses ministérielles, correspondances, avis, prévisions et décisions. L'ensemble de ces documents administratifs sont des archives publiques.

La circulaire du 1er ministre du 2 novembre 2001 précise que la fonction de gestion des archives intermédiaires dans les administrations centrales et les établissements publics de l'Etat doit être assumée par un service ou une cellule spécifique qui doit apparaître clairement dans l'organigramme et être placé à un niveau lui permettant d'exercer efficacement sa mission.

Le référentiel de gestion des archives de la recherche est un document qui précise les durées de conservation et le sort final des documents selon leur typologie avec un renvoi vers le texte de référence. Il est à noter que ce référentiel est actuellement en cours de refonte par le groupe de travail Archives scientifiques de la section des Archivistes des Universités, Rectorats, Organismes de recherche et mouvements étudiants (AURORE) de l'Association des Archivistes Français (AAF).

La loi pour une république numérique du 7 octobre 2016 ou Loi Lemaire pose le principe de l'ouverture des données de la recherche publique. La réutilisation est d'ailleurs encouragée par les autorités européennes et nationales même si toutes

les données publiques ne sont pas réutilisables. Dans le domaine de l'archéologie, par exemple, sont réutilisables non seulement les documents administratifs produits dans le cadre de l'activité d'archéologue mais aussi les photographies, dessins et reproductions d'objets. Seuls les documents sous droits patrimoniaux ne sont pas réutilisables sans l'accord de leur auteur ou de ses ayants-droits.

Le livre 5 du code du patrimoine définit d'ailleurs le patrimoine archéologique comme tous les vestiges, biens et autres traces de l'existence de l'humanité, y compris le contexte dans lequel ils s'inscrivent, dont la sauvegarde et l'étude, notamment par des fouilles ou des découvertes, permettent de retracer le développement de l'histoire de l'humanité et de sa relation avec l'environnement naturel. Il fait également la distinction entre l'archéologie préventive, les fouilles archéologiques programmées et les découvertes fortuites.

La circulaire du 5 juillet 1993 précisent les obligations liées à l'achèvement d'une fouille archéologique préventive et donne notamment des modèles de fiches techniques sur la documentation et les documents finaux de synthèse.

L'arrêté du 16 septembre 2004 définit les normes d'identification, d'inventaire, de classement et de conditionnement de la documentation scientifique et du mobilier issus des diagnostics et fouilles archéologiques. Il précise notamment de quels types de documents est composée cette documentation scientifique.

L'arrêté du 27 septembre 2004 explique comment les rapports d'opérations archéologiques doivent être conçus et présentés.

Les normes de descriptions archivistiques : ISAD(G), ISAAR-CPF, ISDF et ISDIAH définissent les bonnes pratiques de description d'un document, d'un producteur, d'une fonction ou d'une institution de conservation des archives.

La norme générale et internationale de description archivistique (ISAD(G)) décrit le contenu des documents d'archives (fonds ou collections) conservés dans les services d'archives. Elle a pour objectif de décrire du général au particulier, d'adapter les informations au niveau de la description, de faire le lien entre les descriptions et de ne pas répéter les informations. Elle comprend une zone d'identification, une zone du contexte, une zone du contenu et de la structure, une zone des conditions d'accès et d'utilisation, une zone des sources complémentaires, une zone des notes ainsi qu'une zone de contrôle de la description. C'est une norme de référence utilisée pour la rédaction des instruments de recherche.

La norme internationale sur les notices d'autorités utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles (ISAAR CPF) donne les clés pour décrire le producteur des archives de manière structurée. Elle comprend une zone d'identification, une zone de la description, une zone des relations, une zone du contrôle et une zone sur les relations entre les collectivités, les personnes, les familles et les ressources archivistiques. Ainsi, cette norme a pour objectif de décrire les producteurs d'archives et plus largement le contexte des documents.

La norme internationale pour la description des fonctions (ISDF) explique comment décrire celles des collectivités associées à la production et à la gestion des archives. Elle comprend une zone d'identification, une zone du contexte, une zone des relations, une zone du contrôle et une zone sur les relations des fonctions avec des collectivités, des documents d'archives et d'autres ressources. Elle a pour but de normaliser la description des fonctions de ces collectivités dans la rédaction des instruments de recherche.

La norme internationale pour la description des institutions de conservation des archives (ISDIAH) décrit celles-ci de manière normalisée. Elle comprend une zone d'identification, une zone de contact, une zone de description, une zone de l'accès, une zone des services offerts, une zone de contrôle et une zone sur les relations des institutions de conservation avec les documents d'archives et leurs producteurs. Cette norme a pour principal objectif de faciliter la description de ces institutions et de les rendre accessibles au grand public. Il est à noter qu'elle fut très peu utilisée car elle faisait double emploi avec la norme ISAAR CPF.

Étant donné l'ancienneté et la difficile articulation des quatre normes de description archivistique présentées ci-dessus, le Conseil International des Archives réfléchit depuis 2013 à un nouveau standard de description des archives qui les remplacerait. Ce standard intitulé *Records in Context* (RiC) a pour objectif de faciliter l'interopérabilité entre ces différentes normes via le Web de données. Ce standard définit clairement les objets à décrire, leur assignent des attributs et les positionnent les uns par rapport aux autres grâce à un système de relation. Il est composé d'un modèle conceptuel abstrait et global RiC-CM et d'une ontologie OWL, RiC-O, qui transpose le modèle conceptuel en un modèle directement utilisable.

Ressources :

- Code du patrimoine livre I, II et V : https://www.legifrance.gouv.fr/codes/id/LEGITEXT000006074236/ (site web)
- Loi CADA : https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068643/ (site web)
- Circulaire du 1er ministre du 2 novembre 2001 relative à la gestion des archives dans les services et établissements publics de l'Etat : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000774334 (site web)

- Référentiel de gestion des archives de la recherche par la section AURORE de l'AAF, 2009 : https://doranum.fr/stockage-archivage/referentiel-de-gestion-des-archives-de-la-recherche/ (site web)
- Loi pour une république numérique du 7 octobre 2016 (loi Lemaire) : https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/ (site web)
- Circulaire du 5 juillet 1993 sur les obligations liés à l'achèvement d'une fouille archéologique préventive : https://www.culture.gouv.fr/Espace-documentation/Documentation-juridique-textes-officiels/Circulaires-et-instructions-administratives-en-vigueur-relatives-a-l-archeologie
- Arrêté du 16 septembre 2004 portant définition des normes d'identification, d'inventaire, de classement et de conditionnement de la documentation scientifique et du mobilier issu des diagnostics et fouilles archéologiques : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000627559
- Arrêté du 27 septembre 2004 portant définition des normes de contenu et de présentation des rapports d'opérations archéologiques : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000628726
- ISAD(G), norme générale et internationale de description archivistique par l'*International Council on Archives*, 1999 : https://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_FR.pdf
- La description archivistique selon les normes ISAD(G) et ISAAR (CPF) et selon les logiciels libres de gestion des archives par Simon Florentin Adjatan (novembre 2014) : http://www.dan.ilemi.net/IMG/pdf/la_description_archivistique_selon_les_normes_isad.pdf
- ISAAR CPF, norme internationale sur les notices d'autorité utilisées pour les Archives relatives aux collectivités, aux personnes ou aux familles par l'International Council on Archives, 2004 : https://www.ica.org/sites/default/files/CBPS_Guidelines_ISAAR_Second-edition_FR.pdf
- Note d'information DITN/RES/2004/002 sur la diffusion de la norme ISAAR CPF : https://francearchives.fr/fr/file/ad1d24044222f69394059d99c649f5c65eff0029/Note_d%27information_DITN_RES_2004_0
- ISDF, norme internationale pour la description des fonctions (ISDF) par l'*International Council on Archives*, 2007 : https://www.ica.org/sites/default/files/CBPS_2007_Guidelines_ISDF_First-edition_FR.pdf
- ISDIAH, norme internationale pour la description des institutions de conservation des archives par l'*International Council on Archives*, 2008 : https://www.ica.org/sites/default/files/CBPS_2008_Guidelines_ISDIAH_First-edition_FR.pdf
- Note d'information DITN/RES/2008/007 sur la parution des normes ISDF et ISDIAH : https://francearchives.fr/fr/file/444ff42a69792f65b8463e92a00f1f868ba70bb2/Note_DITN_RES_2008_007.pdf
- *Records in Context*, conceptual model par l'*Expert Group on Archival Description* de l'*International Council on Archives*, consultation draft v0.2 (july 2021) : https://www.ica.org/sites/default/files/ric-cm-02_july2021_0.pdf
- (RiC), par l'de l', ICA Congress Seoul, 2016 : https://www.ica.org/sites/default/files/session-7.8-ica-egad-ric-congress2016.pdf, https://www.ica.org/fr/records-in-contexts-ric-les-archives-dans-leur-contexte-une-norme-de-description-archivistique-cre-0
- Sémantisation et visualisation de métadonnées archivistiques : mise en ligne du prototype français PIAAF par Florence Clavaud (mai 2018) : https://www.ica.org/fr/semantisation-et-visualisation-de-metadonnees-archivistiques-mise-en-ligne-du-prototype-francais
- Entrepôt du projet *Records in Context* (RiC) sur GitHub : https://github.com/ICA-EGAD/RiC-O
- Textes réglementaires et durée de conservation (vidéo) par Marie-Laure Bachèlerie-Gouverneur (novembre 2019).
- Traçabilité des activités de recherche et gestion des connaissances - Guide pratique de mise en place (PDF) par Alain Rivet, Marie-Laure Bachèlerie, Auriane Denis-Meyere et Delphine Tisserand.
- Guide méthodologique pour l'archivage des bases de données (PDF) par le CINES, 2004.

2. Conservation à long terme

La préservation à long terme (ou pérennisation) consiste à garantir la sauvegarde et l'accès dans le temps sans perte d'information, en assurant l'intégrité des données, de leur format, voire du support de lecture des fichiers (logiciel et matériel).

Huma-Num a établi un partenariat avec le Centre Informatique National de l'Enseignement Supérieur (CINES) sur l'archivage à long terme des données numériques en sciences humaines et sociales, en collaboration avec le centre de calcul de l'IN2P3. Pour que les données soient éligibles à l'archivage pérenne au CINES, les données doivent être dans un des formats publiés et libres dont la liste est établie par le CINES. L'outil FACILE fournit la liste des formats éligibles ainsi que la possibilité de tester en ligne la conformité d'un fichier pour un archivage pérenne.

Ressources :

- Archivage des données à Huma-Num(vidéo) par Michel Jackobson (2019);
- Sensibilisation à la sécurisation et à la pérennisation des données (vidéo) par Lorène Béchard et Marion Massol (2014).
- Sélectionner les données pour la préservation : enjeux et méthodes (PDF et vidéo) par Magalie Moysan (2020).
- Outil FACILE du CINES : https://facile.cines.fr/

VIII. Partage

Selon le statut des données et la législation en vigueur (cf. VII), le partage porte sur tout ou partie des corpus ou n'est pas autorisé. La mise à disposition de données archéologiques peut être restreinte pour protéger des sites du pillage par exemple. En outre, la diffusion de données à caractère personnel est encadrée par le Règlement général sur la protection des données (RGPD). Dans tous les cas, le partage de données commence par le respect des principes FAIR (cf. II.1.B) et en particulier par l'identification unique et pérenne des ressources : jeux de données, documents, articles…

Ressources :

- RGPD : https://www.cnil.fr/fr/reglement-europeen-protection-donnees

1. Identification pérenne

Un identifiant est une association univoque et stable entre un code alphanumérique et une ressource. Sur le Web, on localise les ressources par leur URL (Uniform Resource Locator), c'est-à-dire l'adresse unique qui permet d'accéder à la ressource en ligne. Une ressource peut être une image, un document, une page web, un fichier de données, une fiche d'une base de données, un concept d'un référentiel, etc. Les identifiants pérennes permettent de référencer les ressources sans ambiguïté, facilitant ainsi leur accessibilité, leur citabilité et leur réutilisation. Les identifiants pérennes sont nombreux et si certains sont proches, chacun a ses particularités. Pour la publication de jeux de données archéologiques dans le Web sémantique, les identifiants pérennes sous forme d'URI (Uniform Resource Identifier) sont utilisés, à la fois pour identifier les ressources (DOI, ARK, Handle), et également pour identifier les concepts issus de référentiels standards (VIAF, GeoNames, PACTOLS). L'identification par l'URI permet un référencement sans ambiguïté et contribue à l'interopérabilité des jeux de données en fiabilisant l'accès permanent aux ressources.

Ressources :

- Identifiants des documents numériques par Jean-Luc Archimbaud (2015)
- Les identifiants pérennes : un aperçu, présentation par l'INIST concernant la production scientifique et les auteurs (2017)

A. Archival Resource Key (ARK)

Archival Resource Key (ARK) est un système d'identifiants mis en place en par la California Digital Library (CDL), pour identifier de manière pérenne des objets de n'importe quelle nature (numérique, physique ou conceptuel). L'attribution des identifiants est effectuée par les autorités nommantes enregistrées, gratuitement, auprès de la CDL qui leur délivre un numéro NAAN (*Name Assigning Authority Number*) servant de préfixe aux identifiants. Chaque autorité nommante doit développer elle-même un système de gestion et de résolution d'identifiants. Par exemple, la Maison de l'Orient et de la Méditerranée, NAAN 76609, a développé le service ARKéo qui permet de créer, gérer et mettre a? disposition des identifiants ARK uniques et pérennes pour les ressources numériques (éventuellement physiques) de ses équipes.

Ressources :

- Page officielle sur les Identifiants ARK
- Forum francophone sur l'identifiant ARK
- Le service ARKEO de la Maison de l'Orient et de la Méditerranée

B. Digital Object Identifier (DOI)

Le *Digital Object Identifier* (DOI), issu du monde des éditeurs et du e-commerce, est fréquemment utilisé pour identifier des articles et des publications en ligne. Il est de plus en plus utilisé pour identifier également d'autres ressources. Le consortium international à but non lucratif DataCite est dédié à l'attribution d'identifiants pérennes DOI pour les données de la recherche. Pour créer un DOI, un enregistrement préalable est nécessaire afin d'obtenir un préfixe (équivalent du NAAN pour ARK). Il peut être obtenu, en France, auprès de l'INIST (CNRS), membre de DataCite, qui est agence d'attribution de l'identifiant pérenne DOI.

Ressources :

- Service d'attribution de DOI par l'INIST : PID OPIDoR

C. Handle

Handle est un système d'identification pérenne géré par la *Corporation for National Research Initiatives* (CNRI). S'il est aussi efficace que ARK, l'utilisation de Handle nécessite de payer pour créer les identifiants. En l'absence de paiement de l'abonnement, l'accès aux données n'est plus effectif et le lien n'est donc plus du tout pérenne. C'est pourquoi nous déconseillons l'utilisation de Handle.

2. Le Web des donnée (Linked Open Data) et le Web sémantique

Le Web des données a pour objectif la publication sur le Web de données structurées et reliées entre elles afin de constituer un réseau global d'informations sous forme de graphe. Le *Linked Open Data* s'y inscrit pour les données ouvertes, libres d'accès, favorisant ainsi leur partage et leur réutilisation. Le Web sémantique fournit un modèle, fondé sur le format RDF pour la mise en œuvre du Web des données, permettant ainsi le partage et la réutilisation des données par les applications et les machines qui accèdent ainsi au graphe de données global.

A. Les triplestores

Les *triplestores* sont des systèmes de gestion de données structurées sous forme de triplets au format RDF. Ils sont aux données RDF ce que les bases de données relationnelles sont aux données structurées avec le modèle entité-association conçu par Peter Chen.

Pour stocker et traiter les données RDF, les *triplestores* s'appuient sur la structure (modèle conceptuel) d'un triplet RDF (Fig. 7). Un triplet RDF est constitué par :

- un sujet : l'URI qui référence l'élément décrit.
- un prédicat : une propriété qui permet de décrire un concept ou une instance d'un concept. Ce prédicat est également identifié par une URI.
- un objet : qui est la valeur de la propriété pour le sujet décrit. Il peut être une valeur primitive (chaîne de caractère, valeur numérique, valeur booléenne, …), un concept ou une instance de concept. Dans les deux derniers cas l'objet sera une URI.

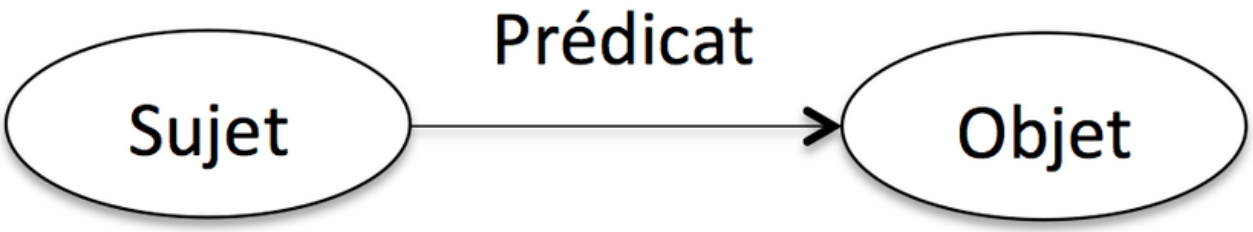


Fig. 7 : Structure d'un triplet RDF.

Voici quelques exemples d'informations qu'il est possible de formaliser en triplets :

- Howard Carter (sujet) a pour métier (prédicat) archéologue (objet)
- Howard Carter (sujet) a pour nationalité (prédicat) britannique (objet)
- Howard Carter (sujet) a découvert (prédicat) le masque funéraire (objet)
- Le masque funéraire (sujet) a appartenu (prédicat) à Toutânkhamon (objet)
- Toutânkhamon (sujet) était (prédicat) un pharaon (objet)...

Voici ce que peut donner le premier exemple formalisé au format RDF utilisant des URI :

`https://en.wikipedia.org/wiki/Howard_Carterhttp://www.cidoc-crm.org/cidoc-crm/CRMsoc/has_role https://en.wikipedia.org/wiki/Archaeology`

Pour stocker les données RDF et les manipuler, il existe plusieurs représentations physiques implémentées dans les *triplestores*. Par exemple :

- la représentation verticale qui utilise une table relationnelle avec trois attributs : sujet, prédicat et objet ;
- la représentation horizontale qui utilise une table relationnelle pour chaque classe du modèle conceptuel et fait correspondre à chaque prédicat de la classe un attribut dans la table ;
- la représentation binaire (Table propriété) qui utilise une table relationnelle pour chaque propriété et deux attributs par table (sujet et objet) ;
- la représentation à base de graphe, plus récente, utilise la structure du graphe RDF ;
- la représentation à base d'autres structures physiques de données (Index, arbre, …).

Les données RDF sont présentent sous deux formes :

- dans des *triplestores* ;
- sous la forme de texte dans un format bien précis qui peut être du XML, on parle alors de RDF-XML (.xml, .rdf, .owl, .rdfs), une liste de triplets au format NTriple (.nt) ou encore une forme contractée de liste de triplets appelé Turtle (.ttl).

Ressources :

- Ali, W.; Saleem, M.; Yao, B.; Hogan, A.; Ngonga Ngomo, A. Storage, Indexing, Query Processing, and Benchmarking in Centralized and Distributed RDF Engines: A Survey. Preprints 2020, 2020050360 (doi: 10.20944/preprints202005.0360.v3).
- https://www.w3.org/TR/rdf-syntax-grammar/
- https://www.w3.org/TR/turtle/
- https://www.w3.org/TeamSubmission/n3/
- https://json-ld.org/

B. OpenArchaeo

Les travaux du Consortium MASA pour l'interopérabilité des données archéologiques se concrétisent par la plateforme OpenArchaeo. L'objectif d'OpenArchaeo est double. Il s'agit d'une part de publier des jeux de données archéologiques dans le Web sémantique, à l'aide d'un *triplestore* MASA dont le modèle conceptuel est fondé sur l'ontologie du CIDOC CRM (cf. VI.4.A), d'autre part de proposer une interface intuitive pour l'interrogation croisée des jeux de données publiés dans OpenArchaeo. L'interface développée par SPARNA propose de construire des requêtes graphiquement à partir de concept archéologiques et les traduit automatiquement en langage SPARQL (langage de requête du Web sémantique) pour interroger les *triplestores*. Inspirée du moteur de recherche ResearchSpace mis en place par le British Museum, le composant SPARNATURAL est applicable à des champs thématiques comme l'archéologie avec OpenArchaeo mais il est aussi en cours de déploiement pour la Bibliothèque Nationale de France et les Archives Nationales.

L'utilisation de l'ontologie du CIDOC CRM étant totalement compatible avec l'infrastructure européenne ARIADNEplus qui a fait le choix d'un modèle conceptuel (AO-CAT) avec une granularité moins fine pour l'accès aux jeux de données. MASA a aligné les deux modèles de telle sorte que les jeux de données d'OpenArchaeo soit également dans la plateforme ARIADNEplus.

Ressources :

- Plateforme Web sémantique OpenArchaeo : http://openarchaeo.huma-num.fr/
- Plateforme Web sémantique ARIADNEplus : https://portal.ariadne-infrastructure.eu/

3. Signalement des données avec ISIDORE

ISIDORE est un moteur de recherche créé en 2010 par la très grande infrastructure de recherche (IR*) Huma-Num en 2010 pour accéder aux données numériques et numérisées de la recherche française et internationale en sciences humaines et sociales (publications, données numériques, profils de chercheurs…). Il permet de rechercher parmi 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 parmi des signalements d'événements (séminaires, colloques, etc.).

ISIDORE collecte les métadonnées par le moissonnage de trois formats ouverts : OAI-PMH, RSS et atom, RDFa. Elles sont ensuite converties an RDF et enrichies par traitements sémantiques des termes de thésaurus scientifiques thématiques ou généraux. Le thésaurus PACTOLS (cf. VI.1.B.a) est exploité par ISIDORE et les jeux de données

d’OpenArchaeo sont indexés dans ISIDORE.

Ressources :

- portail ISIDORE : https://isidore.science/
- Documentation sur ISIDORE : https://documentation.huma-num.fr/isidore/

4. Géolocalisation de l’information avec ArkeoGIS

ArkeoGIS est une plateforme construite par des archéologues et des géographes en vallée du Rhin supérieur afin de partager des informations transfrontalières spatialisées sur le passé. Au cours de son évolution, l’outil s’est ouvert à d’autres chronologies et caractérisations en fonction des besoins des utilisateurs.

L’objectif est de partager et signaler simplement (à partir d’un tableur et de métadonnées https://arkeogis.org/manuel/importation/) des informations sur les sites, le mobilier, les analyses Sur cette base, ArkeoGIS propose d’interroger et de trier les données afin d’obtenir des états (sous forme de tableurs) pour préparer une fouille ou une publication, ou encore des catalogues (*legacy data*, inventaires de masters et de thèses, bases produites dans le cadre de projets...).

Les données déposées dans ArkeoGIS sont alignées avec le thésaurus PACTOLS et préparées pour une intégration dans OpenArchaeo.

Ressources :

- Manuel d’utilisation d’ArkeoGIS : https://arkeogis.org/manuel/

5. Exposition des données avec Nakala

Nakala est un service d’Huma-Num ouvert en 2014 pour partager, publier et valoriser tous types de données numériques documentées (fichiers textes, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes FAIR. Nakala s’appuie sur le *framework*Symfony constitué de trois plateformes : une interface web permettant l’éditorialisation des données, une API et un entrepôt de données. Les données sont stockées au CINES (cf. VII .B) et disposent d’un identifiant pérenne (DOI, cf. VIII.1.B).

Nakala permet de mettre à disposition les métadonnées sur le Web de données et en OAI-PMH.

A. Envoi des données via l’API « Lokala »

L’API « Lokala », développée à la MOM par Hélène Jamet permet d’envoyer les données dans Nakala sous la forme de lots et non pas à l’unité comme avec l’interface native de Nakala.

Cette API permet trois actions :

- rechercher dans les données dont l’utilisateur est propriétaire ou bien sur lesquelles il dispose de droits ;
- envoyer des données avec leurs métadonnées dans Nakala ;
- modifier le contenu des métadonnées ou les données elles-mêmes si elles s’avèrent incorrectes.

Un tableur formaté permet de créer les champs de métadonnées que l’on souhaite ainsi que le contenu à envoyer dans ces champs. Lokala permet donc de rechercher, envoyer et intervenir sur un lot de données publié dans Nakala.

B. Diffusion des URL pérennes

Nakala fournit automatiquement un identifiant DOI (cf. VIII.1.B) à chaque donnée déposée. Elles deviennent « citables » de manière pérenne et peuvent ainsi être liées à des sites de valorisation comme des portails web de diffusion, des bibliothèques numériques, ou bien associées à des publications numériques ou imprimées. Cela permet d'utiliser des images et des textes pour les afficher sur un portail web par l'intermédiaire de l'Api de Nakala.

Il est possible d'accéder à ses données par un lien constitué de la manière suivante : `https://nakala.fr/*identifiant_de_la_ressource*` par exemple `https://nakala.fr/10.34847/nkl.a8eb90mr`

Ressources :

- Accès à Nakala :https://nakala.fr
- Prise en main de Nakala (Romain Boissat)
- Documentation de Nakala : https://documentation.human-num.fr/nakala
- Bac à sable : https://test.nakala.fr,https://demo.nakala.fr
- API Nakala : https://api.nakala.fr/oai2?verb=Identify
- Documentation API Nakala : https://gitlab.huma-num.fr/huma-num-public/notebook-api-nakala/-/blob/master/presentation-api.ipynb
- Présentation par Bruno Morandière lors des journées MASA à Aix en 2018 : https://masa.hypotheses.org/files/2018/12/JourneesMASA-Nakala.pdf
- GitLab de Mickael Nauge présentant des outils pour l'appropriation de Nakala : https://gitlab.huma-num.fr/mnauge/nakalapyconnect
- Lokala application d'import d'images par lots dans Nakala par le service informatique de la MOM : https://lokala.mom.fr/

IX. Réutilisation

Le dernier des principes FAIR, permettre la réutilisation des données, est l'ultime objectif de la science ouverte. Il implique que la personne souhaitant réutiliser des données interopérables sache ce qu'il est autorisé à en faire, donc que le fournisseur de données ait mis en place une politique de licence claire. En outre, pour qu'un jeu de données soit réutilisable, il est indispensable que le (ré)utilisateur puisse avoir confiance dans la qualité de ce jeu de données. Lors de la conférence CIDOC 2018 à Héraklion, Franco Niccolucci (VAST-LAB, PIN - *University of Florence*) a dirigé la session *Heritage data-centric research: are FAIR data fair enough?*, proposant de compléter les « R » de *Re-usable* par *Reliability*(fiabilité) et *Relevance*(pertinence). Des métadonnées précises et complètes, doivent permettre d'évaluer la fiabilité des données (leurs conditions d'acquisition, par exemple) et leur pertinence (le cadre de recherche pour lequel le corpus a été constitué).

La réutilisation de données implique de citer ses sources. En effet, un des freins au partage des données est la quasi absence de valorisation de la constitution et la mise à disposition de corpus dans l'évaluation des chercheurs, contrairement aux publications. Toutefois, d'une part cette situation change très rapidement avec l'engagement des institutions dans la science ouverte, d'autre part une prise de conscience se fait peu à peu que le fait que rien ne protège mieux les travaux, dont la constitution de corpus, que leurs publications citables et référençables. C'est la fonction de l'identification pérenne par DOI proposée par Nakala par exemple (cf. VIII.5.B)

Quand cela est possible légalement, toute publication de données devrait donner accès aux sources (l'enregistrement de terrain par exemple) pour que les traitements et analyses effectuées pour constituer le corpus puissent être éventuellement réitérés et critiqués. Le contenu d'une base de données de fouille archéologique peut présenter de multiples différences avec les fiches d'enregistrement papier effectué sur le terrain : donner l'accès aux scans de ces enregistrements initiaux peut se révéler crucial pour comprendre un jeu de données et en faciliter la réutilisation.

Ressources :

- Les enjeux de la citation (anglais) : https://www.rd-alliance.org/groups/sharing-rewards-and-credit-sharc-ig

1. Licences

Plusieurs licences sont disponibles pour les dépôts de données :

Etalab : https://www.etalab.gouv.fr/licence-ouverte-open-licence

ODbL (Open Database License) : https://opendatacommons.org/licenses/odbl/1-0/

Creative Commons : https://creativecommons.org/licenses/?lang=fr-FR

Ces licences ne sont pas équivalentes. Leur sélection dépend d'une part de la protection souhaitée, d'autre part des choix de l'institution de référence (le ministère de la Culture par exemple privilégie Etalab).

Il faut veiller à ne pas être trop restrictif dans le choix d'une licence, les données partagées doivent pouvoir être réutilisées. Il faut donc autoriser la modification (implémentation) du jeu de données sans nécessairement interdire l'utilisation commerciale (par exemple pour une carte de répartition des sites mentionnés dans un catalogue de musée payant), on préfèrera donc une licence CC-BY (*Creative Commons* avec Attribution) à une CC-BY-NC-SA (*Creative Commons* avec Attribution sans utilisation commerciale et partage aux mêmes conditions) par exemple.

Ressources :

- Documentation officielle Etalab (PDF) : https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf
- Dominus Carnufex : Le libre à la française, ces licences injustement méconnues (HTML, 2015)
- Outils juridiques pour l'*Open Data* (anglais) : https://opendatacommons.org/licenses/by/1-0/
- Véronique Ginouvès et Isabelle Gras : La diffusion numérique des données en SHS - Guide des bonnes pratiques éthiques et juridiques (2018)
- Consortium CAHIER : Guides du groupe « Droits et questions juridiques »

2. Publications scientifiques

A. Data paper

Un *data paper* est une publication décrivant un jeu de données scientifiques via des métadonnées riches, indépendamment des analyses que l'on peut en faire et des résultats que l'on peut en tirer. Il est publié, comme pour un article de revue, après un examen par un comité éditorial. Les *data papers* sont en général regroupés dans des *datas journals*. La publication des *data papers* suit les recommandations des principes FAIR (cf. II.1.B).

Ressources :

Rédiger et publier un data paper (Coopérer en Information Scientifique et Technique)

Publier mes données comme un article scientifique (DORANum)

Liste des *Data journals* disponibles : http://proj.badc.rl.ac.uk/preparde/blog/DataJournalsList

B. Publications en ligne

Les standards de stockage et d'échange des publications numériques sont organisés autour des technologies XML. Le XML est utilisé pour structurer les textes en respectant la norme TEI (*Text Encoding Initiative*), aujourd'hui massivement

utilisé en France et à l'étranger. Les infrastructures de recherche OpenEdition et Métopes proposent des schémas TEI dédiés à l'édition scientifique.

Une difficulté réside dans l'effort d'adaptation de la TEI à l'archéologie qui mobilise souvent une grande quantité de documentation associée à la publication. En outre, les interfaces de lecture existantes proposent peu de solutions d'accès aux images, plans, schémas, etc. indispensables en archéologie.

Le projet Archean (MRSH de Caen), soutenu par le consortium MASA, propose un modèle adapté aux textes archéologiques et aux données mobilisées, ainsi qu'une méthodologie pour articuler les données planimétriques, les données et les textes scientifiques reposant sur des principes d'identification. Ce modèle articule les standards TEI, pour les textes, et EAD (cf. IV.2.C), pour les inventaires de données. Archean a également pour objectif de proposer à la communauté l'outillage nécessaire pour diffuser les textes et la documentation. Il s'appuie sur l'éditeur XMLMind XML Editor (gratuit), avec un environnement dédié à l'archéologie, et le moteur d'affichage XML MaX qui supporte l'EAD et la TEI.

Ressources :

- Projet Archean : https://www.unicaen.fr/recherche/mrsh/document_numerique/projets/archean

C. Publication logiciste

Le programme logiciste initié dans les années 70 par Jean-Claude Gardin propose des solutions pour permettre la lecture rapide des publications et pour clarifier les mécanismes de raisonnement pratiqués dans les constructions archéologiques. Il est fondé sur le constat de l'augmentation exponentielle de la quantité de publications qui rend impossible d'espérer lire l'intégralité de la production. La consultation est devenue la norme et de ce point de vue, le numérique ne fait qu'aggraver les choses. Il s'agit donc de trouver des solutions pour permettre à la communauté de prendre connaissance rapidement du contenu des publications en un temps minimal mais avec la plus grande précision possible.

Les technologies XML et du Web sémantique permettent aujourd'hui de développer et de mettre en œuvre des outils et des publications selon les principes du logicisme dans les meilleures conditions possibles. Le langage XML et la TEI permettent de structurer les textes scientifiques et d'annoter les propositions logicistes qui constituent les unités de base des raisonnements tandis que le modèle d'argumentation CRMinf du CIDOC (cf. VI.4.A) permet d'expliciter les relations entre ces briques logiques. À partir de ces instances XML TEI il est possible de produire automatiquement les diagrammes logicistes adaptés à la lecture rapide : le lecteur pouvant très vite consulter le raisonnement de l'auteur, des faits observés aux conclusions scientifiques.

Un environnement d'édition structuré pour l'éditeur XMLMind XML Editor a été développé pour permettre la structuration des textes, l'annotation des propositions logicistes et leurs mises en relation. Une chaîne de traitement permet ensuite la production des diagrammes logicistes à partir des textes encodés. L'ensemble des documents produits, textes et diagrammes, sont ensuite intégrables dans une solution de publication basée sur le moteur d'affichage MaX. La publication de la fouille du centre paroissial de Rigny (Indre-et-Loire), sur laquelle l'ensemble de ces travaux s'est appuyé, est éditée en ligne par les Presses universitaires de Caen. Les PUC éditent également la revue Arkeotek Journal, selon les mêmes principes d'édition.

Les diagrammes logicistes de ces publications ont été élaborés à partir de texte. Une autre approche consiste à permettre à l'auteur de construire son diagramme logiciste, c'est-à-dire une représentation de son raisonnement, avant la rédaction d'un texte de synthèse le cas échéant, à la manière d'une forme avancée de plan détaillé. C'est dans cette logique qu'a été imaginé Logicist Writer, un outil de conception de diagramme logiciste générant automatiquement le fichier XML associé. Il propose à l'auteur de construire ses propositions et le réseau logique auquel elles appartiennent puis d'exporter le résultat de son travail en XML TEI pour servir de base à sa publication.

Ressources :

- Buard, P.-Y., Zadora-Rio, Chauveau J., Roger J., Marlet O. : Publishing an Archaeological Excavation Report in a Logicist Workflow
- Gardin, J.-C., Une archéologie théorique, Hachette (coll. L'Esprit critique), Paris, 1979.
- Gardin, J.-C., Archaeological Constructs. An Aspect of Theoretical Archaeology, Cambridge University Press, 1980.

- Marlet O., Zadora-Rio E., Buard P.-Y., Markhoff B., Rodier X.: The Archaeological Excavation Report of Rigny: An Example of an Interoperable Logicist Publication

3. Gestion des images avec IIIF

A. Qu'est-ce que IIIF ?

Il s'agit de l'acronyme pour : *International Image Interoperability Framework* (IIIF) qui désigne à la fois une communauté et un ensemble de spécifications techniques dont l'objectif est de définir un cadre d'interopérabilité pour la diffusion et l'échange d'images numériques haute résolution sur le Web. L'initiative IIIF est portée et animée par depuis 2015 par un consortium international constitué de bibliothèques nationales, de musées, d'universités ou instituts de recherche, de portails et agrégateurs généralistes ou spécialisés. La diffusion standardisée permet de rendre consultables, manipulables et annotables les images par n'importe quelle application ou logiciel compatible. Le but est de proposer une expérience utilisateur enrichie en termes d'accès, de manipulation et d'exploitation des images en ligne. Le cadre technique repose sur un ensemble de spécifications qui définissent des API ou services web élaborés de manière concertée au sein de la communauté IIIF.

Ressources :

- : https://iiif.io/
- Spécifications officielles stables : https://iiif.io/technical-details/#stable-specifications

B. Service IIIF360

Depuis 2013, Biblissima joue un rôle moteur dans l'adoption et la promotion de IIIF en France et porte le Consortium IIIF360. Il s'agit d'une offre d'expertise autour de IIIF portée conjointement par Biblissima, le Campus Condorcet et l'IR* Huma-Num. Elle s'adresse à tout projet de recherche ou de diffusion/valorisation de ressources patrimoniales ou pédagogiques mobilisant des images fixes (ou autres types de média supportés par IIIF).

L'offre de services consiste essentiellement en :

- accompagnement de projet et conseil sur-mesures'appuyant sur l'expertise de Biblissima : formations, aide technique ou méthodologique à l'implémentation des APIs IIIF, aide au choix et à la configuration d'outils compatibles IIIF (serveurs d'images, serveurs d'annotation, visualiseurs etc.), assistance à la rédaction de cahier des charges ou au contrôle qualité, aide à la conversion de formats, etc. ;
- stockage de données et de l'hébergement web par l'infrastructure d'Huma-Num ;
- développement informatique.

Ressource :

- Communiqué sur IIIF360 : IIIF360 : Une offre d'expertise autour de IIIF
- Présentation des services IIIF par Biblissima : https://iiif.biblissima.fr/#services-iiif
- Comprendre IIIF et l'interopérabilité des bibliothèques numériques (Régis Robineau)
- Vidéos et supports de présentations du séminaire IIIF360 (24 mars 2021).
- Visualisation et fouille des données : IIIF (Régis Robineau). Cours de Master 2 « Médiation Numérique de la Culture et des Patrimoines » (École Supérieure en Intelligence des Patrimoines, CESR, Université de Tours, 26

novembre et 7 décembre 2020)

- Adopter et utiliser les standards IIIF pour vos corpus d'images numériques (Régis Robineau). Atelier en ligne organisé dans le cadre du colloque « #dhnord2020 - La mesure des images : approches computationnelles en histoire et théorie des arts » (MESHS, 18 novembre 2020)
- Introduction aux protocoles IIIF (Régis Robineau). Présentation faite dans le cadre de la formation au Diplôme de conservateur de bibliothèque de l'Enssib (Villeurbanne, 23 janvier 2019).
- Supports de présentations de la journée « Innover pour redécouvrir le patrimoine écrit », premier événement francophone autour de IIIF (15 mars 2018).
- IIIF en 5 minutes : comprendre les avantages et apports de IIIF du point de vue des institutions et des utilisateurs finaux (Régis Robineau).
- Introduction à IIIF, sur le site de documentation de Biblissima.
- IIIF Frequently Asked Questions (FAQs) (anglais)
- An Introduction to IIIF (Tom Crane, mars 2017)
- Awesome IIIF (liste de ressources utiles maintenue par la communauté IIIF)
- Chaîne YouTube de IIIF

X. Où se former ?

La constitution d'un inventaire exhaustif des ressources disponibles pour se former aux bonnes pratiques numériques est une mission désespérée. En revanche, l'état des lieux proposé ici peut être complété par des offres de formations.

1. Offre de formation du consortium MASA

En complément de son écosystème numérique et afin de favoriser la diffusion des outils et des bonnes pratiques auprès de la communauté archéologique, le consortium MASA propose une série de formations. Elles sont proposées périodiquement et peuvent également être organisées, à la demande, pour un groupe constitué.

Ressources :

- Liste des formations du consortium MASA : https://masa.hypotheses.org/formations

2. Doranum (INIST)

L'Institut de l'Information Scientifique et Technique du CNRS met à disposition une plateforme de formation à distance : "Données de la Recherche : Apprentissage Numérique" (DoRANum), réalisée en partenariat avec le réseau des Urfist. À travers des ressources variées (vidéos, tutoriaux, fiches synthétiques, quiz, etc.), cette plateforme permet de se former à la gestion des données de la recherche en complète autonomie. DoRANum permet également de compléter ces formations à distance avec des dispositifs de formations en présentiel, des ateliers ou webinaires.

Ressources :

- Plateforme DoRANum : https://doranum.fr/

3. URFIST

Le réseau des Unités Régionales de Formation à l'Information Scientifique et Technique inter-académique propose des formations dont l'objectif est de développer au sein de l'enseignement supérieur et de la recherche les compétences en matière d'usage et de maîtrise de l'information scientifique. Les sessions sont régionales et gratuites pour les personnels de l'enseignement supérieur.

Ressources :

- Catalogue des formations URFIST et inscription : https://sygefor.reseau-urfist.fr/#/
- Urfistinfo, le Blog des URFIST : https://urfistinfo.hypotheses.org/
- Site du GIS des URFIST : http://gis-reseau-urfist.fr/

XI. Prolongements

Fig. 8 : L'écosystème du consortium MASA

L'ensemble des bonnes pratiques préconisées dans ce Livre Blanc est articulé autour du cycle de vie des données (Fig. 8). Ce Livre Blanc peut être complété utilement par la lecture d'autres guides de bonnes pratiques, notamment celui sur la gestion des données de la Recherche (HTML ou PDF) par le groupe de travail inter-réseaux « Atelier Données » ou ceux d'autres consortium de l'IR* Huma-Num : celui spécifique aux données 3D en Sciences Humaines et Sociales du consortium 3D-SHS (PDF) ou encore ceux publiés par le consortium CAHIER.