



Atelier Digit_Hum 2024

Peut-on contrôler les vocabulaires ?
Ontologies en SHS :
enjeux, défis, perspectives, outils

Jeudi 3 octobre 2024 – 9h15-17h

École normale supérieure | Salle Dussane

Créer les conditions de l'accès et de l'interopérabilité des données

est devenu depuis plusieurs années une préoccupation majeure, voire une injonction, pour les chercheurs et ingénieurs qui dialoguent sur la méthodologie à mettre en œuvre dans les projets de recherche. La journée d'étude de 2024 sera l'occasion de revenir sur une question transversale à toutes les disciplines SHS : celle de la modélisation des connaissances, à travers les vocabulaires contrôlés, les référentiels, les standards, les thésaurus et plus largement les ontologies.

En rendant compréhensibles, de façon univoque et universelle, l'identification et la caractérisation des objets d'étude, les ontologies s'inscrivent dans le web de données pour structurer et partager des corpus numériques. Est-ce pour autant une solution simple, adoptée par tous ? Jusqu'où faut-il aller pour rendre interopérables les données de la recherche ? Doit-on systématiquement contrôler les vocabulaires et se contraindre à utiliser des listes d'autorité, parfois lacunaires ou peu flexibles ? Il s'agira d'interroger différentes institutions et structures, et de proposer à des scientifiques et des ingénieurs de témoigner de leurs choix passés et actuels, de leurs réticences et de leurs questionnements face à ces standards parfois perçus comme des carcans.

PROGRAMME

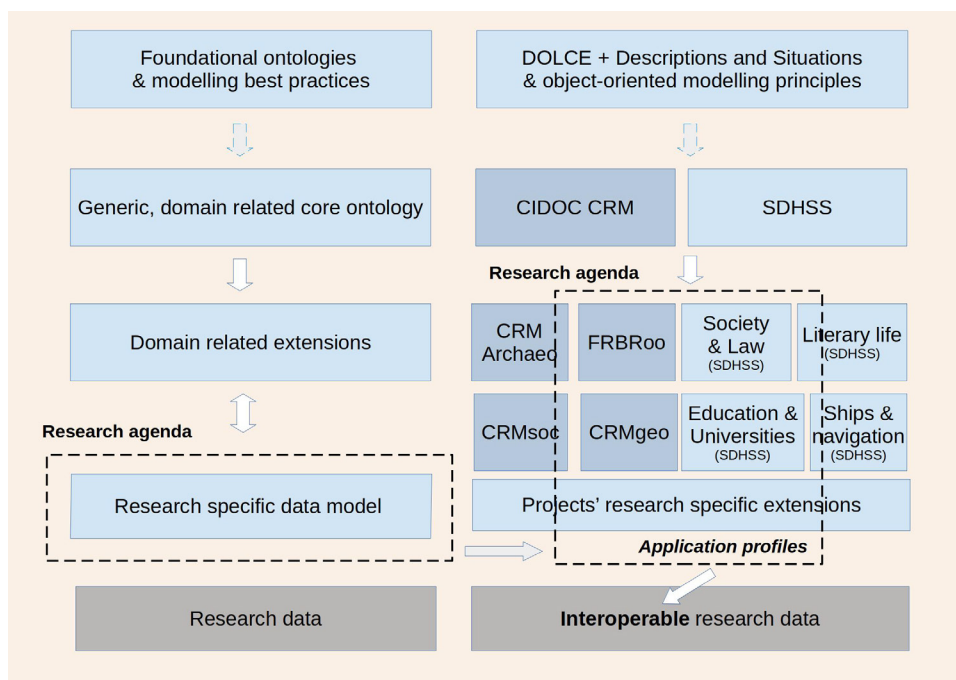
9h15 | Accueil autour d'un café

Mots d'introduction, par **Marie-Laure Massot** (CAPHES, ENS PSL) & **Agnès Tricoche** (AOROC, ENS PSL) : Les dix ans de Digit_Hum !

9h30 | *Semantic Data for Humanities and Social Sciences (SDHSS): une méthode et un écosystème d'ontologies favorisant l'interopérabilité et la réutilisation des données de la recherche en SHS*

Francesco Beretta (LARHRA)

Si les modèles de langage (LLM) suscitent actuellement un engouement, notamment de la part du public, du point de vue de la recherche en SHS il ne faut pas oublier l'importance des graphes de connaissances et des données ouvertes liées (LOD) qui révolutionnent la manière dont l'information peut être partagée et réutilisée à grande échelle, en application des principes FAIR. En mars 2023, le Knowledge Graph de Google aurait atteint « 800 milliards de faits sur 8 milliards d'entités » (Wikipedia, 30/03/2023). Mais quelles sont les conditions pour que les chercheur-e-s en SHS puissent partager les informations produites quotidiennement, en utilisant les mêmes technologies sémantiques, et relier leurs données aux métadonnées des bibliothèques, des archives et des musées, afin de créer un graphe géant d'informations ? Peut-on développer des ontologies partagées, c'est-à-dire des modèles abstraits du monde permettant l'interopérabilité et la réutilisation des données ? La diversité des recherches n'est-elle pas un obstacle insurmontable à ce projet ?



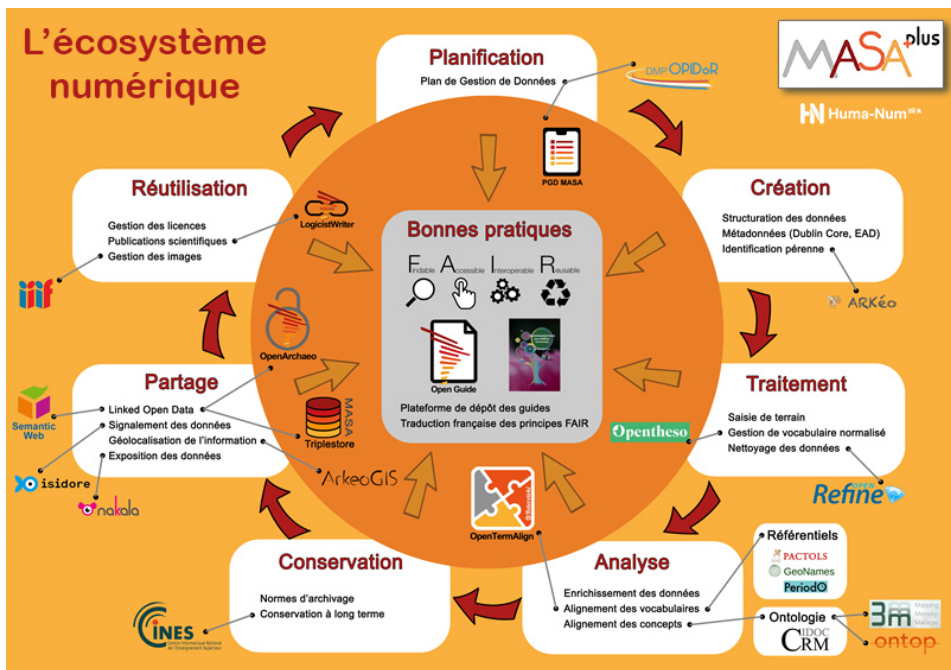
Dans mon intervention, je présenterai la méthodologie adoptée dans le projet Semantic Data for Humanities and Social Sciences (SDHSS) afin d'étendre et d'intégrer le CIDOC CRM en créant un écosystème d'ontologies transdisciplinaire, ouvert et affiné, permettant de documenter sans ambiguïté la signification des informations partagées et de faciliter leur réutilisation pour de nouvelles recherches. Je proposerai également quelques principes permettant de modéliser les différents aspects de la vie sociale, objet d'étude commun en SHS, et de montrer comment, à partir de cette analyse, il est possible d'intégrer les données issues de différents projets dans une même conceptualisation partagée, à l'exemple de la plateforme multi-projet Geovistory.

10h | Présentation des outils du consortium MASApplus pour l'interopérabilité des données

Bruno Baudoin (Centre Camille Jullian)

Le consortium MASA avait mis en place un écosystème numérique selon le cycle de vie des données et conforme aux principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable). Cet écosystème est repris et poursuivi par MASApplus, il fait appel à la fois à des outils open source préexistants, des outils dont le consortium soutient fortement le développement et également MASApplus réalise des développements propres. Désambigüiser les informations, aligner des vocabulaires, utiliser une ontologie servant de couche d'interopérabilité pour faire communiquer des

données hétérogènes ou encore publier des données géo-référencées, dans tous les cas, le consortium veille au respect des standards nationaux et internationaux, en application des principes FAIR et propose un certain nombre de formations à ses outils.



10h30 | Pause

11h | Table-ronde animée par Régis Witz (MISHA)

Avec la participation de :

- **Luc Bouiller** (C2RMF). *PARCOURS, Un modèle sémantique pour l'interopérabilité des données en conservation-restauration des Œuvres du patrimoine.*
- **Kelly Christensen** (École nationale des chartes). *Entre entité et vocabulaire contrôlé, la modélisation d'une base de données collaborative de chansons de geste : LostMa.*
- **Jean-Pierre Girard** (UMR Archéorient/MOM). *D'HyperThesau à SIAMOIS, le partage et l'alignement de vocabulaires de recherche contrôlés, au cœur d'une vision de partage et d'interopérabilité des données archéologiques.*
- **Florian Hivert** (MSH Val de Loire/Huma-Num). *L'ontologie du CIDOC-CRM, standard pour modéliser les données patrimoniales.*

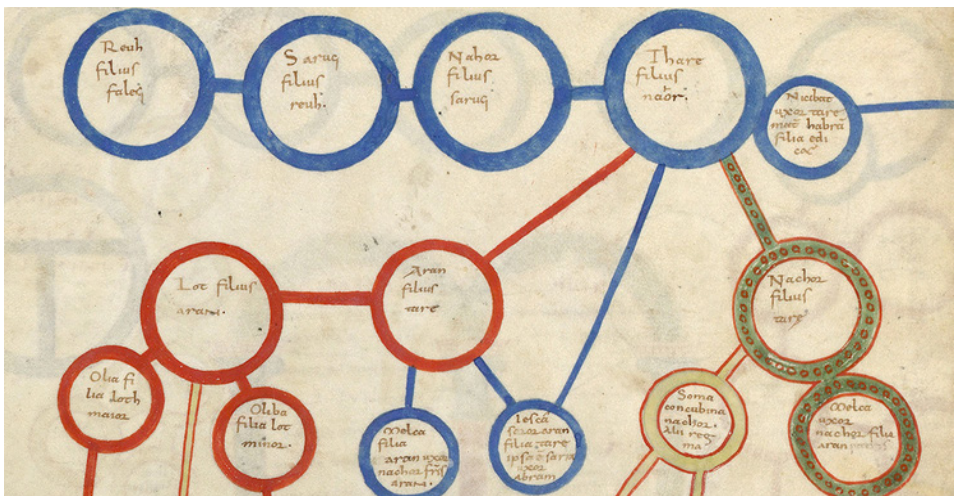
- **Alexey Lavrentev** (consortium ARIANE, UMR IHRIM/ENS de Lyon). *Thésaurus de typologie textuelle, entre les théories littéraires et la linguistique de corpus.*
- **Guillaume Porte** (UMR ARCHE - Université de Strasbourg). *ESTRADES, plateforme de soutien à la recherche autour de l'édition numérique de corpus.*
- **Florence Weber** (Centre Maurice Halbwachs/ ENS PSL). *ArchEthno : créer des thesaurus adaptés aux savoirs de l'ethnographe pour construire des bases de données localisées et pluridisciplinaires.*

12h30 | Buffet en Rotonde

13h30 | Les référentiels Biblissima et leur rôle pour l'interopérabilité et l'ouverture des données

Emmanuelle Morlock (Hisoma, Biblissima+) et **Eduard Frunzeanu** (Biblissima+)

L'interopérabilité des données occupe une place centrale dans l'écosystème du projet Biblissima depuis ses débuts, avec pour objectif la création d'un accès unifié aux données de la recherche sur la transmission des textes depuis l'Antiquité à l'époque moderne. Pour ce faire, plusieurs obstacles sont à surmonter : la diversité des modèles, la plupart du temps idoine à chaque base partenaire, adoptés pour gérer les données ; la diversité des formats utilisés pour leur structuration (SQL, XML-EAD, XML-TEI, JSON) ; la multiplicité des graphies retenues pour libeller une même entité. Le format XML pivot défini afin d'établir des équivalences entre les modèles des bases de données et les référentiels mis en place rendent effective l'interopérabilité dans le portail Biblissima, tout en restant suffisamment flexibles pour permettre l'intégration de nouveaux types de ressources. Notre intervention mettra en lumière les avantages et les limites des solutions de gestion retenues.



14h-17h | Ateliers pratiques

1. Régis Witz (MISHA), Dublin Core, etc. : *par où commencer pour structurer ses données ?* (20 minutes)

Cette initiation propose un rappel de la notion de métadonnée et des principes de base d'utilisation des ontologies, afin de rendre son projet plus FAIR.

2. Blandine Nouvel (Centre Camille Jullian) et Miled Rousset (Maison de l'Orient et de la Méditerranée), *Créer et publier un thésaurus avec Opentheso* (40 minutes)

Cet atelier propose une méthodologie et des conseils pratiques pour la création d'un thésaurus qui s'appuie sur l'utilisation du gestionnaire de thésaurus Opentheso, outil normalisé, libre et gratuit. Les exemples et les recommandations d'organisation et de gestion quotidienne du vocabulaire seront issus de l'expérience de re-construction du thésaurus Pactols pour l'archéologie. L'objectif est de sensibiliser les participants aux bonnes pratiques conduisant à produire un vocabulaire contrôlé le plus FAIR possible.



3. Florian Hivert (MSH Val de Loire/Huma-Num), *Un outil de MASApplus, le CIDOC-CRM* (40 minutes)

L'ontologie CIDOC CRM est un standard depuis quelques dizaines d'années pour modéliser les données patrimoniales (musées, bibliothèques, archives, archéologie) et permettre de les rendre interopérables. Mais, avec une centaine d'entités et le double de propriétés, l'apprentissage de l'ontologie reste complexe et nécessite un fort investissement pour s'approprier son fonctionnement global et ses logiques. Un jeu de cartes a été imaginé pour manipuler les entités et les propriétés du CIDOC et réfléchir au mapping d'une manière ludique.



4. Vincent Alamercury (ENS de Lyon/LARHRA), *Modéliser, produire, explorer, publier des données FAIR : OntoME et Geovistory, un environnement virtuel de recherche pour les SHS* (40 minutes)



L'environnement virtuel de recherche Geovistory est un écosystème libre d'applications permettant aux projets de recherche en SHS, tout comme aux chercheurs individuels, de produire et de publier des données de façon collaborative et cumulative dans le respect des principes FAIR. En son cœur, la Toolbox offre un cadre convivial pour produire et explorer ses données. Chaque projet peut personnaliser son modèle en choisissant des profils - sélections cohérentes de classes et de propriétés - appliqués à son domaine de recherche. Les profils sont élaborés et mis à disposition de la communauté dans OntoME, environnement de gestion d'ontologies, qui permet de construire collaborativement des modèles de données dans le cadre conceptuel du CIDOC CRM et de SDHSS. Geovistory inclut aussi des points d'accès SPARQL accompagnés des outils d'exploration de graphe Sparklis et OntoExplorer. Enfin, un module de publication permet de facilement déployer un site web en utilisant des webcomponents adaptés.



Geovistory

Organisation de l'atelier Digit_Hum 2024 :

Marie-Laure Massot (CAPHÉS, UAR 3610, CNRS-ENS)

Agnès Tricoche (AOROC, UMR 8546, CNRS-ENS-EPHE)

En collaboration avec Régis Witz (MISHA)

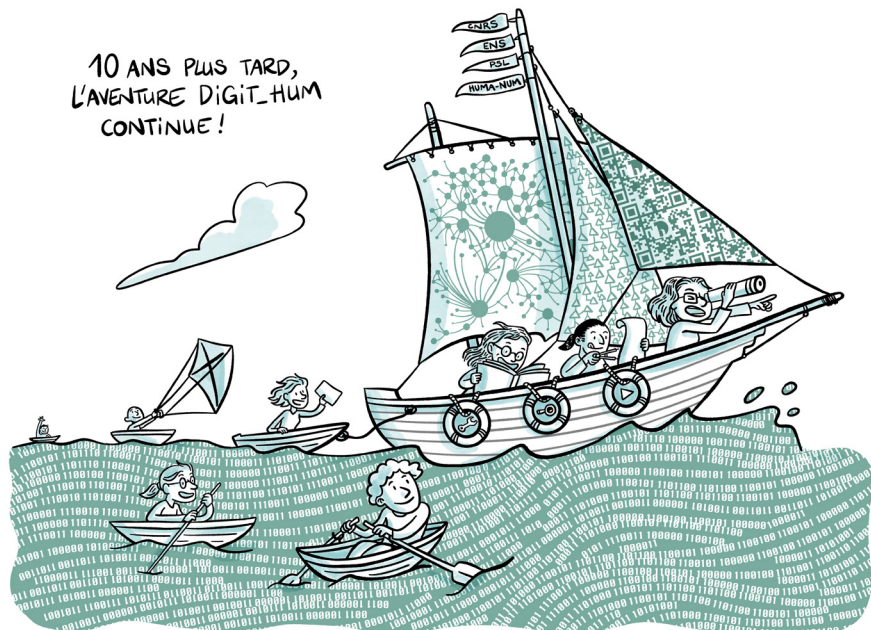
EXPOSITION « DIGIT_HUM : 10 ANS ! »

Initiée par trois ingénieur.e.s en 2014, à l'École normale supérieure de Paris, l'initiative Digit_Hum constitue une archive des humanités numériques en train de se faire. À l'occasion des 10 ans de Digit_Hum, cette exposition revisite une décennie d'échanges, d'enrichissements et d'acculturation par le prisme des illustrations réalisées par les différentes artistes qui ont illustré les ateliers organisés chaque année.

A partir du 19 septembre 2024

Ecole normale supérieure, 45 rue d'Ulm

Escalier du hall d'entrée



Morgane Parisi

Notes

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

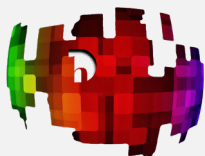
.....

.....

.....

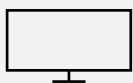
.....

Notes



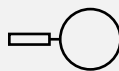
DIGIT_HUM

Humanités numériques
entre savoirs et savoir-faire



Vidéos

des ateliers
Digit_Hum



Enquête

sur les humanités
numériques à l'ENS



Portraits

d'acteurs en humanités
numériques



Ressources

pour s'informer,
se former, s'orienter

et un **portfolio** d'artistes dessinant les humanités numériques

Retrouvez toute notre activité sur
digitum.huma-num.fr

Digit_Hum est une initiative visant à échanger sur les différents usages qui sont faits des humanités numériques et à structurer ces nouvelles pratiques à l'échelle de l'École normale supérieure et de l'université Paris Sciences & Lettres.

Cette initiative est portée par le CAPHÉS (Centre d'Archives en Philosophie, Histoire et Édition des Sciences),
et AOROC (Archéologie et Philologie d'Orient et d'Occident).

Elle bénéficie du soutien de l'EUR Translitteræ et de l'Observatoire des humanités numériques de l'ENS.