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

Emmanuelle Morlock, Eduard Frunzeanu

▶ **To cite this version:**

Emmanuelle Morlock, Eduard Frunzeanu. Les référentiels Biblissima et leur rôle pour l'interopérabilité et l'ouverture des données. Atelier Digit_Hum 2024, Marie-Laure Massot (CAPHES CNRS/ENS PSL – UAR 3610); Agnès Tricoche (AOROC CNRS/ENS/EPHE PSL – UMR 8546); Régis Witz (CNRS, UAR 3227 MISHA), Oct 2024, Paris, France. ⟨hal-04774384⟩

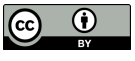
HAL Id: hal-04774384

https://hal.science/hal-04774384v1

Submitted on 8 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

Atelier Digit_Hum 2024

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

Jeudi 3 octobre 2024 – 9h15-17h

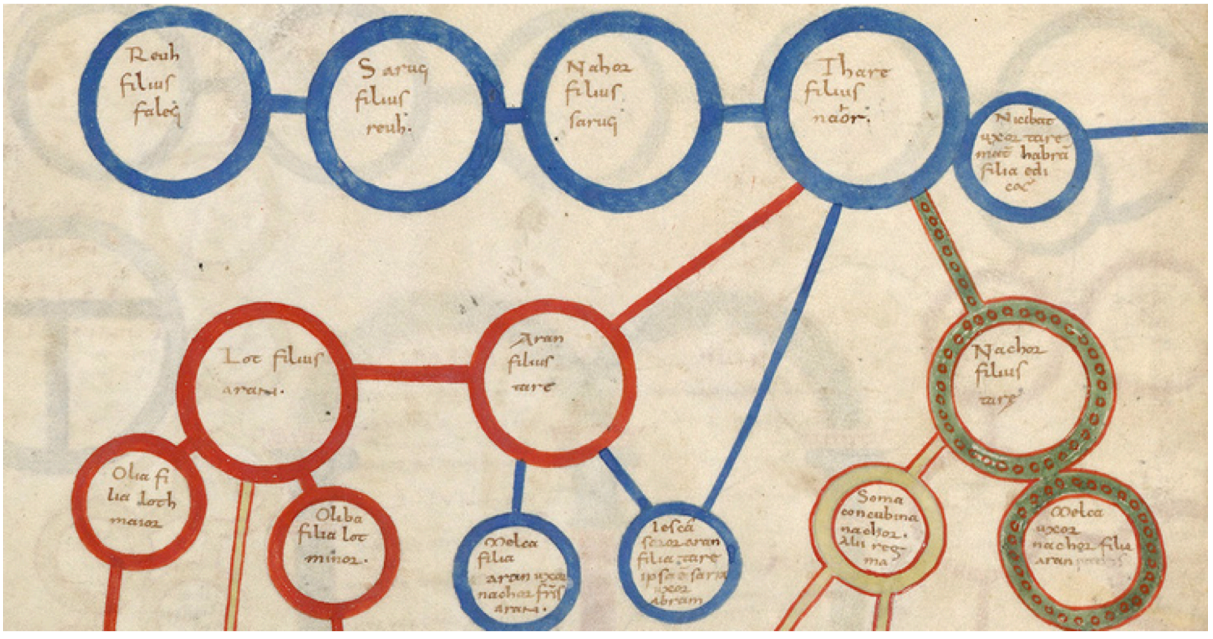
École normale supérieure | Salle Dussane

https://digithum.huma-num.fr/atelier/2024/

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 idoines 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.



[Attention ! Texte partiel : introduction, démonstration et conclusion]

Slide 1

Introduction 2-3 min

Le projet Biblissima s'est construit dans son ambition et ses méthodes sur l'objectif de donner accès à des données scientifiques et patrimoniales dont la complémentarité et la présence numérique étaient évidentes, mais qui n'étaient pas exploitables de manière intégrée du fait de l'hétérogénéité des systèmes, des outils et des modélisations employés pour les produire.

L'interopérabilité occupe depuis le départ une place centrale dans Biblissima et cette place n'a fait que se renforcer depuis la conception du projet en 2011.

A l'opposé de la notion de « carcan » évoquée dans l'argumentaire de cette journée, cette mise en interopérabilité procède d'une structuration sémantique des données et de leurs relations qui s'effectue a posteriori, sans impacter l'autonomie des bases sources ou leurs besoins de modélisation spécifiques.

Cette démarche est en effet mise en œuvre à partir de grandes campagnes d'intégration dont on va expliquer la logique dans la suite de l'exposé. On verra qu'elle ne peut cependant pas se passer de référentiels normalisés pour agréger une documentation scientifique d'origines multiples et offrir des points d'entrées unifiés sur les manuscrits et imprimés, les collections anciennes, les acteurs, les lieux, les enluminures, etc. du patrimoine écrit ancien.

Cette présentation a pour objectif d'aider à mieux comprendre « de l'intérieur » la manière dont l'interopérabilité est mise en œuvre dans le portail biblissima.

Notre exposé se déroulera en 3 temps :

- Je commencerai par une présentation rapide du projet global pour permettre de prendre la mesure de son périmètre et de son ambition

- Puis Eduard expliquera les principes et le fonctionnement de la chaîne de traitement de l'interopérabilité, en faisant le point sur ses limites et ses apports

- Je reprendrai la parole pour une illustration plus concrète de la manière dont les outils peuvent être exploités facilement pour améliorer l'interopérabilité, par d'autres projets, même complètement extérieurs à Biblissima.

Slide 4

Un rapide panorama

Biblissima construit et développe depuis maintenant plus de 12 ans un observatoire des cultures écrites anciennes.

Que vous connaissez probablement déjà… Le laboratoire AOROC, fait en effet partie des 17 équipes scientifiques fondatrices et parties prenantes de son comité de direction. La directrice scientifique et technique de Biblissima, Anne Marie Turcan Verkerk est également membre d'AOROC.

D'un point de vue institutionnel, Biblissima est porté par le Campus Condorcet. Il a été sélectionné à deux reprises dans les appels à projets d'équipements structurants pour la recherche du programme d'investissements d'avenir de l'État français (repris aujourd'hui dans le plan France 2030), et est financé jusqu'en 2029.

De 2012 à 2021, le premier ÉquipEx Biblissima a eu pour objectif créer un un accès unique et simple à la documentation savante sur l'histoire de la transmission des textes du moyen-âge et de la renaissance.

L'infrastructure numérique de Biblissima a été organisée dès lors autour de 4 réalisations :

- Le portail de découverte des données
- La plateforme de référentiels permettant de gérer les entités de référence autour desquelles les ressources du portail peuvent être interconnectées (personnes, lieux, œuvres, cotes de documents…)
- Le moteur de recherche moissonnant les bibliothèques numériques de manuscrits et imprimés anciens conformes au standard de diffusion des images sur le web IIIF
- Un catalogue de toutes les ressources numériques financées à un titre ou l'autre par les financements equipex. Ils se répartissent en 5 grandes catégories : les bibliothèques numériques, les catalogues ou répertoires, les bases de données scientifiques, les corpus spécialisés, les éditions de textes et pour finir les logiciels ou services numériques

Le seconde phase de financement, intitulée ÉquipEx "plus", a ensuite démarré en octobre 2021.

Son programme est centré d'une part sur la consolidation du socle technique et de l'autre sur l'extension du périmètre des données.

En effet, le périmètre temporel de Biblissima+ intègre désormais toute la période antique.
Le périmètre typologique s'ouvre également à de nouveaux supports porteurs d'écrits comme les sceaux ou les inscriptions, avec de vrai défis comme le traitement de toutes les langues anciennes ou les écritures sans polices unicode.
L'ajout de la bibliographie scientifique disponible en ligne dans des plateformes comme Persée, Hal voire Istex, avec des liens directs sur le texte intégral, va encore plus enrichir l'information accessible.

Pour l'après 2029, les tutelles se sont engagées à pérenniser les emplois de l'équipe technique et une labellisation par le Ministère de l'enseignement supérieur et de la recherche est en préparation.

Slide 5

Voici quelques chiffres qui représentent l'état actuel des données dans l'infrastructure.

Slide 6

L'élargissement concerne également le degré d'intégration des communautés scientifiques dans le projet et son rôle structurant, autour de 7 pôles de compétence, appelés "clusters".

Ils réunissent les chercheurs, les étudiants, les ingénieurs, les conservateurs, et tous les professionnels impliqués dans Biblissima+.

Les thématiques ont été définies à partir des grands domaines d'innovation investis par les équipes fondatrices pour mener leurs recherches dans le domaine des cultures écrites anciennes.

Les questions d'appropriation, d'accompagnement, de modélisation de référentiels sont très présentes dans les travaux des clusters.

Slide 7

On peut résumer les grands objectifs principaux actuels de Biblissima par trois points :

- La création et le développement d'un portail et de référentiels sur les cultures écrites anciennes, au niveau national, reste central
- Vu l'ampleur du périmètre, il ne faut cesser d'agréger de nouveaux jeux de données. Cela se fait dans plusieurs directions : les livrables du programme de l'ÉquipEx, les résultat de l'appel à projet annuel, ou encore les collectes et moissonnages de ressources libres réalisées par l'équipe
- Enfin, la structuration de communautés autour de pôles d'innovation, regroupant les membre de biblissima par thématique mais ouverts à tous les volontaires

Ce panorama ainsi posé, je laisse la parole à Eduard qui va vous présenter tout ce qui permet de comprendre la chaîne de traitement des ressources agrégées dans le portail.

Eduard : La gestion de l'interopérabilité dans le portail Biblissima

Slide 33

Pour terminer sur une note plus concrète, nous vous présentons maintenant une méthode relativement simple et accessible à tous, permettant d'améliorer l'interopérabilité de jeux de données liés au patrimoine écrit ancien, même s'ils sont complètement extérieurs à Biblissima.

Slide 34

Pour cette illustration, sous forme de déroulé d'étapes, notre choix s'est porté sur le cas particulier des éditions en XML TEI.

Il y a deux motivations principales à ce choix :

- D'abord, c'est un exemple développé récemment pour un tutoriel présenté en août dernier au Congrès Digital Humanities 2024 à Washington.
- Ensuite on a choisi l'exemple des éditions électroniques car il y un enjeu particulier pour leur intégration. En effet, elles n'ont pas vocation à être intégrées sous forme de texte intégral, conformément à la logique du portail. Mais il est possible de créer des liens croisés via les entités nommées. Cela est valable pour n'importe quelle édition publiée en TEI qui dispose d'url stables pour ses pages.

Slide 35

Le processus global comporte 5 étapes.

On importe le fichier XML/TEI dans l'outil de traitement afin d'extraire les noms de personnes et de lieux avec leurs identifiants.
On normalise éventuellement les formes récupérées pour de meilleurs résultats puis on lance la réconciliation.
La réconciliation consiste pour openrefine qui va se connecter à l'API dédiée de data.biblissima et lancer un algorithme d'appariement qui retournera des résultats.
Après la validation de ces résultats, on n'aura plus qu'à exporter les alignements dans un format TEI pour pouvoir insérer le résultat par un banal copier/coller.

OpenRefine est un logiciel libre qui s'installe en local sur un ordinateur et avec lequel on interagit via son navigateur web. C'est d'abord un outil de nettoyage et de manipulation de données assez puissant mais très facile à utiliser grâce à une interface graphique qui présente les données sous la forme familière de tableaux.
OpenRefine sait interagir avec des API et offre un langage de scripts permettant toutes sorte d'usages plus avancés.

Je vais maintenant dérouler ce même processus avec une série de capture d'écran qui vous permettront de comprendre la logique du processus.

Slide 36

On part d'un fichier TEI dans lequel des noms de personnes et de lieux sont balisés à l'aide de l'élément name.

Slide 37

A l'import du fichier XML on clique sur la première occurrence de la balise qui nous intéresse, en l'occurrence <name>.

On obtient un tableau avec trois colonnes :

- Le nom « duc Jean Ier de Berry
- Le type « place » ou « person »
- L'identifiant « maison »

Slide 38

Après l'import, on a souvent une étape de nettoyage et de normalisation des données afin d'obtenir de meilleurs résultats aux étapes suivantes.

Slide 39

Comme on a à la fois des noms de lieux et de personnes, il faut procéder en deux temps. Ici on a d'abord filtré sur le type "personne".
On lance la réconciliation sur la colonne « name » via le menu. Puis on choisit le service de réconciliation. Ici on choisit bien entendu le service « Wikibase Biblissima » que l'on a enregistré au préalable pour qu'il s'affiche dans la liste.

Slide 40

A cette étape, on vérifie que l'appariement va bien se faire sur une entité de type « être humain », qui correspond à la propriété Q168.

Slide 41

La réponse de l'API s'affiche dans la colonne, la barre de progression donne une représentation de la proportion de cellules appariées automatiquement. On aussi des suggestions quand cela n'a pas été possible.

Slide 42

Généralement on fait plusieurs itérations de réconciliation en précisant la requête afin d'obtenir les meilleurs résultats possibles. Ensuite il faut valider les alignements.

On peut exploiter les liens directs vers Biblissima pour vérifier ou corriger les propositions.

Slide 43
On récupère l'identifiant Biblissima dans une colonne.

Slide 44
On ouvre une interface de patrons de code pour définir l'export TEI qui est une arborescence XML

Slide 45
Ici on a choisi de générer une liste de personnes
Slide 46
L'export est intégré par copier / coller dans l'en-tête du fichier TEI

Slide 47
Un aperçu de la manière dont l'édition pourrait s'afficher dans le Portail

Slide 48
Une exploitation dans l'utilisation TEI Publisher sur le fichier d'exemple, pour lequel on a aussi traité les noms de lieux et extrait les latitude longitude de biblissima pour permettre la visualisation sur une carte

Slide 49
A partir de cette base, de multiples variations sont possibles, avec des traitements d'import / export plus complexes. La réconciliation peut aussi être utilisée pour améliorer la qualité.

Slide 50
A vous de jouer maintenant ! Nous avons mis à disposition le référentiel des descripteurs iconographiques et proposons plusieurs tutoriels pour sur la réconciliation, avec ou sans openrefine.

Slide 51
En résumé, on a vu en détail la méthode de mise en interopérabilité de Biblissima. Elle est a posteriori, ce qui combine le besoin de normalisation pour l'interconnexion et l'automatisation, et les besoins scientifiques de modélisations spécifiques des données sources. La plateforme mise en place par Biblissima gère des identifiants de référence basés sur le même système que Wikidata. C'est une plateforme ouverte aux contributions dans les deux sens : pour l'enrichissement et pour l'exploitation.

Plus elle s’enrichit et grossit, plus les opérations de dédoublonnage sont importantes et lentes. On essaie de travailler avec les équipes sur l’amont, notamment pour le dédoublonnage et les exports.
L'accompagnement et la formation au numérique font partie des missions de Biblissima, on a fait un gros effort sur les référentiels et la réconciliation depuis cette année.
Le nouveau chantier qui démarre actuellement porte sur la question des mises à jour : on a lancé deux études pour essayer de déterminer ce qui peut être automatisé et comment.

Merci pour votre attention.