Enrichissement sémantique de données d'archives sonores d'ethnomusicologie par alignement.

Nedra Mellouli, Aude Julien Da cruz lima

Enrichissement sémantique de données d'archives sonores d'ethnomusicologie par alignement

Nédra Mellouli*, Aude Julien Da Cruz Lima **

*LIASD(EA 4383), Université Paris8 Vincennes, Saint Denis
n.mellouli@iut.univ-paris8.fr,
**LESC (UMR 7186), CNRS, Université Paris Nanterre
aude.da-cruz-lima@cnrs.fr

Résumé. ATMAH : "Alignement Tool for Music Archive heritage" est un projet multidisciplinaire financé en 2016 par la COMUE [1] de l'Université Paris Lumières. Ses disciplines phares sont principalement le web des données et les sciences de l'information appliquées à la SHS. Son objectif principal est d'associer les partenaires culturels (BnF, Musée du quai branly, LESC et la Maison des Cultures du Monde) impliqués dans la diffusion d'archives sonores en ethnomusicologie, afin d'améliorer la gestion des vocabulaires d'indexation selon trois aspects : accès, enrichissement et interopérabilité. Il est donc fondamental de renforcer leur valorisation au niveau national et international dans le contexte des données ouvertes liées. Le défi de l'étude proposée est de démontrer l'intérêt de l'interconnexion et la faisabilité de sa réalisation entre plusieurs bases de données web et systèmes de plates-formes, avec des vocabulaires aussi bien internes spécifiques (structurés en SQL, MARC, etc.) qu'externes ("hub" de dimension internationale, comme langue française du consortium Bnf RAMEAU et MIMO de Europeana, conformes aux standards du web des données). A partir des cas d'utilisation dérivés des classifications d'instruments de musique et d'autres vocabulaires du domaine (tels que la description vocale, les supports audio, les genres musicaux traditionnels, les entités nommées, etc.), il sera nécessaire d'établir des scénarios d'utilisation et des spécifications à tester en fonction des besoins des professionnels de l'information et de la recherche (échange, alignement semi-automatisé, visualisation graphique et dynamique des données, annotation collaborative). Dans la première étape du projet, nous avons ciblé notre étude sur l'amélioration de la gestion des vocabulaires contrôlés dans la plateforme audio web des archives sonores utilisée par l'équipe du CREM et sur les outils et procédures d'alignement. Au delà des archives du CREM, l'outil visé par le projet ATMAH est étroitement lié aux besoins en cours des partenaires (nationaux et internationaux) en tant que consortiums (TGIR Huma-Num du CNRS, France), LabEx Past in Present cluster, Paris Lumières Université, Europeana Sounds.

1. ATMAH : Alignement Tool for Music Archive heritage s'inscrit dans le cadre de l'appel à projets COMUE 2016 de l'Université Paris Lumières

1 Vers une interopérabilité des vocabulaires d'indexation dans l'environnement numérique et du web

La diffusion et l'échange d'information dans un environnement numérique en ligne nécessite de partager des vocabulaires communs pour améliorer la mise en correspondance (interconnexion) des données tout en respectant la spécificité des terminologies métiers. A ce jour où des demandes de partage d'information entre des communautés nouvelles émergent, la plateforme d'archives sonores (Telemeta) [2] utilisée par le CREM ne permet pas seule de lever ce verrou technologique dès lors que les vocabulaires métiers restent spécifiques et dépendants des ressources de chaque équipe. Pour exploiter le potentiel de l'environnement numérique du web, il faut mettre en place un mécanisme d'interopérabilité sémantique à partir de vocabulaires pivots et d'outils de gestion et d'alignements permettant de définir les liaisons appropriées aux différents termes du vocabulaire de référence. Pour résoudre ces problèmes, ce travail contribue à l'étude des possibilités de représenter et de gérer ces vocabulaires dans des formats RDF (SKOS et OWL) permettant la construction de ressources liées à un niveau sémantique, et la recherche de solutions afin de les connecter à la plateforme Telemeta. Ensuite, face à la pluralité et la diversité de ces terminologies métier il est nécessaire de penser des solutions permettant d'améliorer l'accès et l'interopérabilité des vocabulaires contrôlés de la plateforme d'archives sonores du CREM afin de 1) renforcer les passerelles avec des vocabulaires externes de référence, 2) évaluer la stabilité et la pérennité des ressources, 3) valoriser les ressources des producteurs dans le contexte du Linked Open Data. Enfin ce travail vise à dynamiser et étendre les réseaux scientifiques aux niveaux national et international, notamment à travers l'élaboration de passerelles multilingues.

2 Méthodologie et application

Durant la première partie du projet qui s'est tenue en 2016 et 2017, nous avons structuré l'étude en deux phases. Lors de la première phase nous avons procédé à l'analyse des différents vocabulaires dont nous disposons et à établir les schémas de mapping les plus pertinents. Les formats SKOS et RDF ont été retenus pour la représentation des vocabulaires sources. Pour ce faire, nous avons utilisé OpenRefine(Mathieu, 2018) pour nettoyer, transformer le vocabulaire d'un format à un autre et l'étendre aux services web et aux données externes. Une fois le vocabulaire formaté, la seconde phase vient répondre à deux questions : 1) comment aligner ce vocabulaire source avec un vocabulaire cible ? 2) Quels vocabulaires cibles sont intéressants quant à l'enrichissement du vocabulaire source ?

RAMEAU [3] et MIMO(Manguinhas et al., 2016) sont deux vocabulaires cibles conformes aux standards du web de données que nous avons sélectionné pour la thématique instruments de musique. Enfin, la tâche d'enrichissement et de mise en correspondance des vocabulaires sources avec des données externes exige la mise en place d'une démarche qualité et d'évaluation d'un certain nombre d'outils d'alignement tels que ONAGUI [4], YAM++(Ngo et Bellah-

2. https://archives.crem-cnrs.fr/
3. http://rameau.bnf.fr/utilisation/liste.htm
4. ttps://github.com/lmazuel/onagui

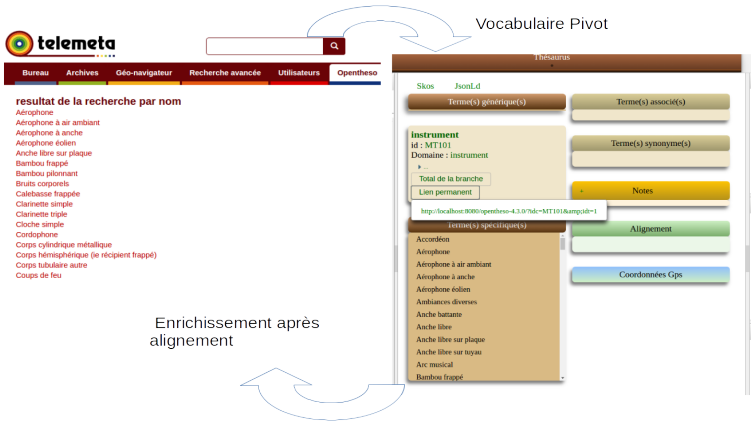

FIG. 1 – *Interface d'enrichissement du vocabulaire dans Telemeta par OpenTheso (Julien Da cruz Lima et Mellouli, 2017)*

sene, 2012), CultuurLink [5] ou OpenTheso (cf.Figure 1). Cette démarche cherche à répondre à trois critères : 1) optimiser le taux de couverture du vocabulaire source par les résultats d'alignement, 2) maximiser le matching positif et minimiser les contradictions, 3) écarter les ligneurs spécifiques. Pour cela, la première étape d'exploration du vocabulaire a été dédiée à l'extraction des termes utilisés par les données de Telemeta et qui concernent les instruments de musique. Cette extraction a permis dans un premier temps de constituer une base de vocabulaires bien identifiée. Lors de la seconde étape, nous avons cherché à enrichir ce vocabulaire par des méthodes d'alignement souvent utilisées pour trouver des correspondances sémantiques entre une ontologie source et une ontologie cible. Notre contribution a consisté à considérer et modéliser une ontologie spécifique aux vocabulaires extraits depuis les données de Telemeta et à l'aligner avec différentes ontologies cibles. Différents tests expérimentaux des procédures d'alignements ont été entrepris sur les archives sonores et audiovisuelles d'ethnomusicologie (du CREM) qui portent sur les référentiels d'instruments de musique et après avoir étudié les outils d'alignements disponibles dans la littérature. Cette étape a été très difficile à réaliser avec satisfaction puisque la plupart des outils d'alignements disponibles avaient été non concluants et souvent très spécifiques au domaine d'application. Néanmoins, nous avons retenu quatres outils qui se distinguent par leur interface graphique et par leur simplicité à savoir OnaGui, Cultuur Link et Yam++, ainsi que le gestionnaire OpenTheso incluant des fonctionnalités d'alignement. Grâce à ces outils nous envisageons de collaborer avec les développeurs afin de les adapter à notre besoin et de pouvoir comparer les résultats que nous obtiendrons sur nos vocabulaires, voire même sur de nouveaux vocabulaires.

5. http://cultuurlink.beeldengeluid.nl/app/

3 Conclusion

Le défi de cette étude et l'objectif de ATMAH est de démontrer l'intérêt de l'interconnexion et la faisabilité de sa réalisation entre plusieurs bases de données web et systèmes de plates-formes, avec des vocabulaires internes spécifiques et externes. A partir des cas d'utilisation dérivés des classifications d'instruments de musique et d'autres vocabulaires du domaine, nous avons expérimenté notre méthodologie sur un volume de données indexées avec les différents vocabulaires contrôlés. Ce volume correspond à environs 75000 fiches documentaires à ce jour avec une progression de 5 à 10 % par an. Les vocabulaires contrôlés utilisés peuvent contenir entre une centaine et plusieurs milliers de termes. En particulier, le vocabulaire des instruments de musique est très complexe. Il peut varier entre les formes génériques et vernaculaires couvrant des centaines de langues du monde entier, avec les diversités des translittérations en fonction des pratiques et des époques. Les résultats tangibles obtenus depuis le début du projet sont principalement d'ordre organisationnelle et méthodologique qui vise à mettre en place sur le long terme un réseau de partenaires interinstitutionnel et pluridisciplinaire et d'une coopération effective exploitant la procédure d'alignement visée pour les données du CREM. Les résultats tangibles visés en perspective sont d'ordre qualitatif avec l'enrichissement des données des différents partenaires et la contribution à la création de référentiels du domaine comme les expressions vocales, l'enrichissement des vocabulaires avec les résultats d'alignement, l'enrichissement sémantique des données indexées avec des référentiels pivots du LOD.

Références

Julien Da cruz Lima, A. et N. Mellouli (2017). Alignment tools for music archive heritage (atmah). *The International Association of Sound and Audiovisual Archives (IASA), Berlin*.

Manguinhas, H. ., V. Charles, A. Isaac, T. Miles, A. . Lima, A. Néroulidis, V. Ginouvès, D. Atsidis, M. Brinkerink, M. Hildebrand, et S. Gordea (2016). Yam++ : A multi-strategy based approach for ontology matching task. *The 16th European Networked Knowledge Organization Systems (NKOS16)*.

Mathieu, S. (2018). Nettoyer et préparer des données avec openrefine. Atelier pour les journées du consortium masa, (https://msaby.gitlab.io/atelier-openrefine-MASA/).

Ngo, D. et Z. Bellahsene (2012). Yam++ : A multi-strategy based approach for ontology matching task. *ten Teije A. et al. (eds) Knowledge Engineering and Knowledge Management. EKAW 2012. Lecture Notes in Computer Science, vol 7603. Springer, Berlin, Heidelberg 14*, 421–425.

Summary

ATMAH is a multidisciplinary study gathering computer science and information science applied to SHS. Its main purpose is to associate french cultural and scientific partners (BnF, Quai branly Museum, LESC, LIASD and the Maison des Cultures du Monde) involved in ethnomusicology audio archive dissemination, to improve the management of indexing vocabularies with regards to three aspects : access, enrichment and interoperability. The question

is how to enhances their valorisation at a national and international level in the context of the Linked Open Data. The challenge of the proposed study is to prove the interest of interconnection and the feasibility of its realization between several web data bases and platform systems, with specific internal vocabularies (structured in SQL, MARC, etc.) and external vocabularies ("hub" of international dimension, as french language of Bnf RAMEAU and MIMO consortium of Europeana, conform to the standards of the data web). Based on use cases derived from classifications of musical instruments, and other vocabularies in the domain (such as voice description, audio carriers, traditional music genres, named entities, etc.) it will be necessary to establish use scenari and specifications to be tested in accordance with the needs of information and research professionals (exchange, semi-automatised alignment, graphical and dynamical datavisualization, collaborative annotation). On a first step of the ATMAH project in 2016-2017, we started to work on both improving controled vocabularies management in Telemeta web audio plateform used by CREM team and testing alignment tools and procedures. ATMAH is a COMUE project [6] closely linked with the partners' ongoing programs (national and internation) as consortiums (TGIR Huma-Num from CNRS, France), LabEx Past in Present cluster, Paris Lumières University, Europeana Sounds.

6. ATMAH : Alignement Tool for Music Archive heritage s'inscrit dans le cadre de L'appel à projets COMUE 2016 de l?Université Paris Lumières