Linking subject labels in Cultural Heritage Metadata to MIMO vocabulary using CultuurLink

Hugo Manguinhas[1], Valentine Charles[1], Antoine Isaac[1], Tom Miles[2], Aude Lima[3], Ariane Néroulidis[4], Véronique Ginouvès[4], Dimitra Atsidis[5], Michiel Hildebrand[6], Maarten Brinkerink[5], Sergiu Gordea[7]

[1]Europeana Foundation, The Hague, The Netherlands
{hugo.manguinhas,valentine.charles,antoine.isaac}@europeana.eu
[2]The British Library, London, United Kingdom
{tom.miles@bl.uk}
[3]The Centre de Recherche en Ethnomusicologie, Paris, France
{aude.da-cruz-lima@mae.u-paris10.fr}
[4]The Maison Méditerranéenne des Sciences de l'Homme, Aix-en-Provence, France
{ariane.neroulidis@gmail.com, veronique.ginouves@univ-amu.fr}
[5]Netherlands Institute for Sound and Vision, Hilversum, The Netherlands
{datsidis@beeldengeluid.nl, mbrinkerink@beeldengeluid.nl}
[6]Spinque B.V., Utrecht, The Netherlands
{michiel@spinque.com}
[7]Austrian Institute of Technology, Vienna, Austria
{sergiu.gordea@ait.ac.at}

Keywords: Vocabulary Alignment, Metadata, Cultural Heritage, Europeana, MIMO, CultuurLink

The Europeana Sounds[1] project aims to increase the amount of cultural audio content in Europeana. It also strongly focuses on enriching the metadata records that are aggregated by Europeana. To provide metadata to Europeana, Data Providers are asked to convert their records from the format and model they use internally to a specific profile of the Europeana Data Model[2] (EDM) for sound resources. These metadata include subjects, which typically use a vocabulary internal to each partner.

The problem is that the values in subject fields come too often as simple literals (strings) that are specific to one (or a couple of) language(s) - the one(s) of the Data Provider. For Europeana to take full advantage of subjects from these vocabularies for purposes such as cross-lingual search, it is essential that they are connected with richer, multilingual data. A first solution to this problem is to semantically enrich metadata for individual cultural objects with links to concepts from a (multilingual) vocabulary (say, 'vocM'). Such new object-vocM links can be used to later provide more semantics and labels in multiple languages for search indexes or display functions. A second option is to perform alignment at the level of vocabularies, linking the elements of an original

[1] http://www.europeanasounds.eu/
[2] http://pro.europeana.eu/page/edm-documentation

(local) vocabulary (say, 'vocL') to semantically related elements from a richer vocabulary, i.e., creating new vocL-vocM links that can be used to enhance the value of existing object-vocL links. Both solutions present many challenges[3] [1,2]. In the cultural sector, more experience needs to be gained, in order to determine their level of feasibility for obtaining 'good enough' results, and answer basic questions as (1) can we find good vocabularies? (2) can we identify suitable processes and tools? (3) how much manual effort is needed and how much can it be automatized?

This paper focuses on exploring the feasibility of vocabulary alignment. This task is often deemed more successful when done manually by domain experts. However, it becomes too labour intensive for large vocabularies. Some tools have proposed a semi-automatic approach to make the task less labour-intensive yet still taking benefit from the user expertise required to assert the right alignments. We conducted an experiment with some Europeana Sounds Data Providers to use of a vocabulary alignment tool, CultuurLink[4], to identify alignments between the subject terms from local vocabularies and a semantically richer target vocabulary.

CultuurLINK[5] is an agile vocabulary alignment tool developed by Spinque. It aids the user in the process of identifying alignments between vocabularies, by seamlessly combining both automatic and manual approaches. It is the successor of the Amalgame framework [3] developed in the EuropeanaConnect project [4].

We asked Data Providers that were contributing data to Europeana as part of WP1 of the Europeana Sounds project to select a collection of metadata records (about sounds recordings, interviews, radio programmes) that could contain musical instruments terms. The data being mapped using the Europeana Sounds EDM profile[6], an extension of the Europeana Data Model, which is itself heavily based on Dublin Core. We considered only musical instruments terms within subject fields (dc:subject) as instructed by collection owners. A total of 6 datasets containing a total of 10,406 metadata records were obtained from the providers and evaluated in this experiment:

- The British Library (BL) participated with 3 collections: A selection of Asian instruments (1,099 records) from the "Colin Huehns Asia Collection"[7]; a selection from the “Peter Cooke Uganda Collection”[8] (1,312 records); and the “Keith Summers English Folk Music Collection”[9] (1,326 records). All three collections were chosen for their rich variety of different musical instruments from each region.

---

[3] See special session of the 13th European NKOS Workshop: https://at-web1.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2014/programme.html

[4] http://cultuurlink.beeldengeluid.nl/app/#

[5] http://2015.semantics.cc/michiel-hildebrand

[6] http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_task-forces/EDMSound//TF_Report_EDM_Profile_Sound_301214.pdf

[7] http://sounds.bl.uk/World-and-traditional-music/Colin-Huehns-Pakistan

[8] http://sounds.bl.uk/World-and-traditional-music/Peter-Cooke-Uganda

[9] http://sounds.bl.uk/World-and-traditional-music/Keith-Summers-Collection

- The Centre de Recherche en Ethnomusicologie (CREM) participated with a test collection[10] of 25 records published in the CD "Musical Instruments of the World" which shows a great variety of traditional instruments with generic terms in french (from the 4 organological families of the SH classification) and corresponding vernacular terms.
- The Maison Méditerranéenne des Sciences de l'Homme (MMSH) participated with a collection of 25 records about folk music.
- The Netherlands Institute of Sound and Vision (NISV) participated with a collection of 6,608 records (not available online) containing commercial 78 rpm records (Handelsplaten) from different genres like light music, classical music and opera.

As a significant number of terms within subject fields of the Europeana Sounds data are related to musical instruments we chose the Musical Instruments Museums Online[11] (MIMO) vocabulary, a reference vocabulary used in a previous Europeana-related project[12], as target vocabulary for our experiment following the recommendations made in [5]. The MIMO vocabulary is a multilingual controlled vocabulary of musical instruments built to ensure consistency of classification for the musical instruments[13]. It is a result of an alignment of a vernacular classification with the professional "Hornbostel-Sachs" classification[14]. The vocabulary has been built with English as pivot language, and translations in seven other languages have been added after.

The goals of our experiments were to:

- evaluate the use of a semi-automatic tool like CultuurLink for a concrete vocabulary alignment case, and;
- assess the coverage of the MIMO vocabulary for enriching Europeana Sounds datasets.

We decided to focus on the vocabulary terms as they are used, i.e. present within the subject fields of the metadata sent to Europeana. We chose to do this, as opposed to aligning the full vocabulary used by the providing institution, since:

- these were not available for use outside the organization and/or in a data structure that suits a vocabulary alignment tool (e.g. SKOS), and furthermore, we did not have the opportunity or the resources to develop an export to SKOS for each vocabulary; and,
- we preferred to report on alignments for the subjects used in the source datasets and not on all possible subjects.

We asked the providing institutions to design and apply alignment strategies in cultuurLink and then evaluate the alignments (i.e., validate the links and assign them a

[10] http://archives.crem-cnrs.fr/archives/collections/CNRSMH_E_1990_014_001/
[11] http://www.mimo-international.com/MIMO/
[12] http://pro.europeana.eu/project/mimo
[13] http://www.mimo-db.eu/InstrumentsKeywords/
[14] http://www.mimo-international.com/documents/Hornbostel%2520Sachs.pdf

type of SKOS mapping link). Once all the participants had finished their task we collected the alignment results and summarized the findings.

In general, the Data Providers found that applying a simple matching technique using just an exact (using equals comparison) string matching of preferred labels on source and target (i.e. the initial strategy we had created), was enough to identify more than 50% (reaching 80% in some cases) of all possible alignments for musical instruments. When using this strategy also incorrect alignments were identified due to polysemy reasons (e.g. "ban" or "zang" which means singing or song was a candidate match to the instrument "zang", a sort of cymbals or clapper bells). They were able to use the tool to discard it by manually confirming the ones they were interested on. All Data Providers also tried more elaborate strategies to discover the possible remaining alignments, typically by using a less restrictive string matching function.

There was a consensus from the Data Providers that the experiment was successful and they were able to understand and work with the vocabulary alignment tool with good level of success.

In our presentation, we will report in more details on the most notable achievements and findings from our experiment.

References

1. Isaac, A., Manguinhas, H., Stiller, J., Charles, V.: Report on Enrichment and Evaluation. The Hague, Netherlands (2015), http://pro.europeana.eu/taskforce/evaluation-and-enrichments.
2. Isaac, A., Manguinhas, H., Charles, V., Stiller, J., et al: Comparative evaluation of semantic enrichments. Technical report (2015). Report available at http://pro.europeana.eu/taskforce/evaluation-and-enrichments. Data archive available at: https://www.assembla.com/spaces/europeana-r-d/documents?folder=58725383
3. Ossenbruggen, J., Hildebrand, M., Boer, V.: Interactive vocabulary alignment. In: Proc. 15th International Conference on Theory and Practice of Digital Libraries, pp. 296-307. ACM (2011). http://semanticweb.cs.vu.nl/lod/tpdl2011/paper.pdf.
4. J. Wielemaker, V. de Boer, A. Isaac, J. van Ossenbruggen, M. Hildebrand, G. Schreiber, S. Hennicke. Semantic workflow tool available. EuropeanaConnect Deliverable D1.3.1. October 2011. http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaConnect/Deliverables/ECONNECT-D1.3.1-Semantic%20Workflow%20Automation%20Method%20Implementation.pdf
5. Isaac, A., Manguinhas, H., Charles, V., Stiller, J., et al: Selecting target datasets for semantic enrichment. Technical report (2015). Report available at http://pro.europeana.eu/taskforce/evaluation-and-enrichments.