Exploring Audiovisual Archives through Aligned Thesauri

Victor de Boer[1,2], Matthias Priem[3], Michiel Hildebrand[4], Nico Verplancke[3], Arjen de Vries[4], and Johan Oomen[1]

[1] Netherlands Institute for Sound and Vision, Hilversum, The Netherlands
{vdboer,joomen}@beeldengeluid.nl
[2] Department of Computer Science, VU University Amsterdam, The Netherlands
[3] Flemish Institute for Archiving, Gent, Belgium
{matthias.priem,nico.verplancke}@viaa.be
[4] Spinque B.V., Utrecht, the Netherlands
{michiel,arjen}@spinque.com

Abstract. As audiovisual archives are digitizing their collections and making these collections available online, the need arises to also establish connections between different collections and to allow for cross-collection search and browsing. Structured vocabularies, made available as Linked Data, can be used as connecting points by aligning thesauri from different institutions. In this paper, we present a case study where partial collections of two audiovisual archives are connected by aligning their thesauri. We report on the conversion of one of the thesauri to SKOS and on the subsequent application of an interactive alignment tool "CultuurLINK". Finally, we introduce an cross-collection browser which uses the produced alignment to allow users to explore connections between the two collections.

Keywords: audiovisual archives, thesaurus alignment, cross-collection browsing

1 Introduction

The task of audiovisual archives is to store audiovisual heritage from various sources and to make this material available to media professionals, researchers and the general public. In recent years, audiovisual archives have initiated large-scale digitisation projects and started ingesting born-digital material. Collections are being made available on the Web so they can be accessed by various user-groups and making them available on the Web. However, these collections are mostly still only available through their own Web interfaces and are rarely connected to outside collections and other information sources.

At the same time, end users are more and more expecting to be able to access, browse and search across different collections. Especially for media researchers, such cross-collection exploration is extremely valuable[1]. Structured vocabularies, published online as SKOS[2] and made available as Linked Data[3] offer excellent opportunities to provide this type of integration.

In this document, we describe a case study where parts of two national audiovisual collections are described using SKOS thesauri. The collections are archived by the Flemish Institute for Archiving (VIAA)[5] and the Netherlands Institute for Sound and Vision (NISV)[6]. Both institutions manage large digital archives, composed of a variety of sources. The material comes from the public broadcaster(s), regional broadcasters and / or cultural heritage institutions. This archive material is accessible to diverse audiences, such as customers themselves, research or education. The thesauri of the two institutions are aligned using an interactive and transparant alignment tool. We finally introduce a cross-collection browser which uses the produced alignment to allow users to explore connections between the two collections. The contributions of this paper are the following:

- We describe the entire pipeline of a real-world, international use case that illustrates the end-user benefit of aligned SKOS thesauri;
- We present a method and tools for converting XML thesauri to SKOS;
- We introduce CultuurLINK, an interactive tool for thesaurus alignment;
- We present an application that enables cross-collection search and browsing using the aligned thesauri.

2 The two institutions and their data

2.1 Netherlands Institute for Sound and Vision and the GTAA

NISV is the largest audiovisual archive in the Netherlands, with more than 800,000 hours of radio, television, film and music archived. The archive makes its collection accessible to diverse audiences, including media professionals, the creative industries, education and the general public. Through research and innovation, the institute has developed into a broad cultural institution that plays a central role through his knowledge and infrastructure within the archive and media sectors. NISV makes its collection available online through various end-user services, including services for the creative industry, education and research.

GTAA The Common Thesaurus for Audiovisual Archives[7] (GTAA). The GTAA is used by NISV to annotate the different collections. The GTAA closely follows the ISO-2788 standard for thesaurus structures and consists of several facets for describing TV programs: subjects, people mentioned, named entities (Corporation names, music bands etc), locations, genres, producers and presenters. The GTAA, available as SKOS, contains approximately 180.000 terms and is actively maintained, being updated as new concepts emerge on television. Approximately $20,000$ terms have broader or narrower relationships. Nine concept schemes divide the thesaurus terms of content in geographical terms, persons, genres, etc. The thesaurus includes about $90,000$ scope notes, and $33,542$ terms are related to each other.

[5] `http://viaa.be`
[6] `http://beeldengeluid.nl`
[7] http://datahub.io/dataset/gemeenschappelijke-thesaurus-audiovisuele-archieven

OpenImages The GTAA thesaurus is used to annotate the entire NISV collection, however, due to licensing issues, only a small part of this collection is made publicly and freely available using Creative Commons licenses. This "open images" collection is a set of $1,700$ video items freely available on the Web, mostly consisting of Dutch public news items from the mid 20^{th} century. The Open Images dataset can be accessed through a web portal[8] and a set of APIs. There also is a version available in the Resource Description Framework (RDF). For the research described here, we use this RDF version.

2.2 VIAA and the VRT thesaurus

VIAA is the Flemish institute for archiving and was founded in 2012. VIAA digitizes, archives and makes available material from more than 80 organizations. Among these organizations are the Flemish public broadcaster VRT, regional broadcasters, archives and cultural heritage institutions. The digitized and archived material is made available to education, research and the public (through public libraries).

Conversion of the VRT thesaurus The VRT archive is managed using a digital system, and since 1986, the media items were already annotated using a central keyword list. Recently (in 2014), this thesaurus was greatly downsized and imported to the media management system. The thesaurus currently consists of 102,172 terms. To enable reuse of the thesaurus and linking within the project, we converted the thesaurus to the SKOS format. First, the thesaurus was exported in XML format, which provided insight into the structure and other features of the thesaurus. Unlike GTAA, the VRT thesaurus does not separate different types of terms in different concept schemes. The 'what', 'where' and 'who' terms are all be found in the same list. Terms have alternative labels and relations to other terms. Each term, moreover, has been given a unique ID, which we can use for the assignment of URIs for the $102,172$ `skos:Concept` instances. The original relations were mapped to SKOS relations as follows:

- 'ParentID' attributes, indicating hierarchical relations are mapped to `skos:broader` and `skos:narrower` relations. Examples of this include geographical part-of relations as well as subclass relations. In total $97,744$ concepts have a broader or narrower relation.
- Terms without parentID indications are modeled as `skos:TopConcepts`. In total there are $4,429$ top concepts.
- Preferred and alternative labels are mapped to `skos:prefLabel` and `skos:altLabel`, respectively. The labels receive RDF language tags specifically indicating the Flemish dialect of Dutch ('`@nl-be`').
- 'Relation' attributes are mapped to `skos:related` RDF triples. Examples include cities and their football club ('Amsterdam' - 'Ajax').
- 'Explanation' attributes are mapped to `skos:scopeNotes`. Only a small fraction of terms have this attribute (212 terms).

[8] `http://openimages.eu`

As, at the time of conversion, the SKOS thesaurus has not yet received an official status and is not yet published by either VIAA or VRT, the URI namespace is left unspecified in the converted SKOS thesaurus[9]. The conversion script is based on NodeJS and uses the Skosify library[10], and is available as open source code [11]. It is well-documented and can be reused to convert similar thesauri or to be re-run in case adaptations or additions are made to the thesaurus in the original management system.

Collection subset Unfortunately, for the Flemish audiovisual data, neither the entire archive, nor any subset are at the moment openly licensed. They are only available for research or educational purposes. Nevertheless, we selected a subset of the VRT video collection of 35,000 items. Like Open Images, this is only a small subset of an archive containing more than 1 million records. These items are contemporary television broadcasts and are annotated with thesaurus terms.

3 Thesaurus Alignment

3.1 CultuurLINK

To align the two SKOS thesauri, we used the functionalities of the CultuurLINK tool [12]. CultuurLINK was based on research on interactive alignment and the Amalgame tool described in [4]. CultuurLINK extends that tool with a more efficient backend implementation and an improved end-used interface. CultuurLINK is an interactive web-based vocabulary alignment tool in which strategies can be constructed to optimally produce correspondences (links) between concepts of two or more thesauri. It features an intuitive end-user interface where collection managers arrange different work-flow elements using a drag-and-drop interface. These elements include different filters and word-matching techniques. The interface also allows for inspection and evaluation of intermediary or final results. This highly interactive alignment allows the collection managers - who know the different features and peculiarities of the source and target thesauri - to develop a specialized and transparent workflow which leads to high quality alignment between the two vocabularies. CultuurLINK features (fuzzy) string matching strategy elements, regular expression operations, and basic Natural Language Processing options (e.g. stemming). Additional strategy elements can be applied to select concepts based on structural properties.

3.2 Aligning the two thesauri

Several strategies have been designed for the alignment of the VRT and GTAA thesaurus corresponding to different types of terms. The reason for this division

[9] We use `http://example.org` as temporary namespace in the produced SKOS files
[10] `https://github.com/NatLibFi/Skosify/`
[11] `https://github.com/viaacode/skoscreator`
[12] `http://cultuurlink.beeldengeluid.nl`

is that different types of terms have different properties. For example, person names consist of at least two parts (In GTAA "Last-Name, First-Name", in the VRT thesaurus "First-Name Last-Name") requiring the use of a regular expression and topic terms might take a singular or plural form, calling for word stemming. For optimal matching a division into four strategies corresponding to different types of terms was made. These categories share some characteristics in terms of the way the actual terms are constructed. Therefore, by isolating them, we can use specific workflows of string matching techniques for each of the term types. There are four such substrategies for:

- **Topics**. Topic terms are generic concepts. (e.g. "transportation").
- **Locations**. Concepts denoting geographical names (e.g. "Amsterdam")
- **Entities**. For example organization names (e.g. football club "Ajax")
- **Persons**. For example: media producers or persons appearing in news footage.

The CultuurLINK screenshot in Figure 1 shows the entire strategy for Topics visually. It shows the different filters to isolate the topics from the two vocabularies. Subsequently, string-matchers and other building blocks are used to identify matching terms. The bottom half of the screen shows the inspection part of the tool, where the user can inspect and evaluate intermediate or final mappings. The strategies can be explored at `http://cultuurlink.beeldengeluid.nl`[13].

Term type	Links
Topics	4,167
Entities	2,197
Locations	4,011
Persons	11,265
Total	21,640

Table 1. Number of established links between the thesauri per strategy

Table 1 shows how many links are eventually found between the thesauri. A total of $21,640$ links are found, which indidates that 21% of the VRT terms are mapped to a corresponding GTAA term. The percentage might seem low, however, the two thesauri each have a different focus (Dutch vs. Flemish). GTAA, for example, contains many names of Dutch media producers or actors who do not appear in any Flemish media items, and therefore absent from the VRT thesaurus. The same holds for geographical terms. For the more neutral 'Topics', the overlap is in fact significant. GTAA includes $4,683$ topics ("subjects") while the VRT thesaurus contains $25,155$ potential subjects (identified by the exclusion of people and geographical concepts). In total, $4,167$ mappings are found, which corresponds to 89% of the GTAA terms. $4,011$ out of $8,617$ locations are matched, corresponding to 47%. Inspection of the unmatched location terms shows that these are mostly smaller places, appearing in one but not in the other thesaurus. A formal analysis of the unmatched terms would give more insight into the exact quality of the alignment and whether it matches expectations.

[13] A strategy can be revisited by entering a session identifier "vrt_onderwerpen","vrt_plaatsen", "vrt_namen" and "vrt_personen"

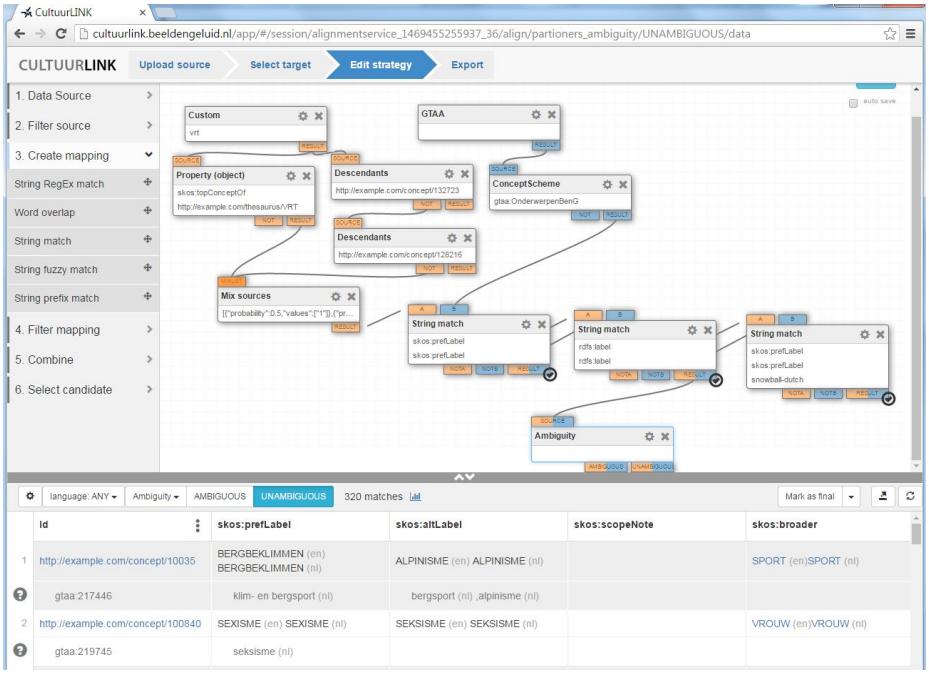

Fig. 1. Screenshot of the CultuurLINK tool, showing the strategy for Topics

The exported links, together with the SKOS thesauri themselves are published on github[14]; they are also accessible through an online triple store[15] allowing for browsing, downloading and querying using the SPARQL protocol.

4 The Demonstrator

4.1 User interface

The demonstrator can be reached at `http://link.spinque.com/VIAA-1.0/`. As part of the collections cannot be made available to the general public, a password is needed[16]. To visualize the functionality of the demonstrator, a screencast of the demonstrator is publicly available at `https://youtu.be/iOJvcHRfvDY`. Figure 2 shows two annotated screenshots of the application.

After logging in, a user starts by using a search term. The interface displays matching results based on titles and descriptions. The concepts with which videos are annotated are presented. For these concepts, the interface shows whether they occur in a single, or in both thesauri (yellow / blue dot). The concepts can be

[14] `http://github.com/biktorrr/gtou_taalunie`
[15] `http://semanticweb.cs.vu.nl/test/`
[16] Available upon request

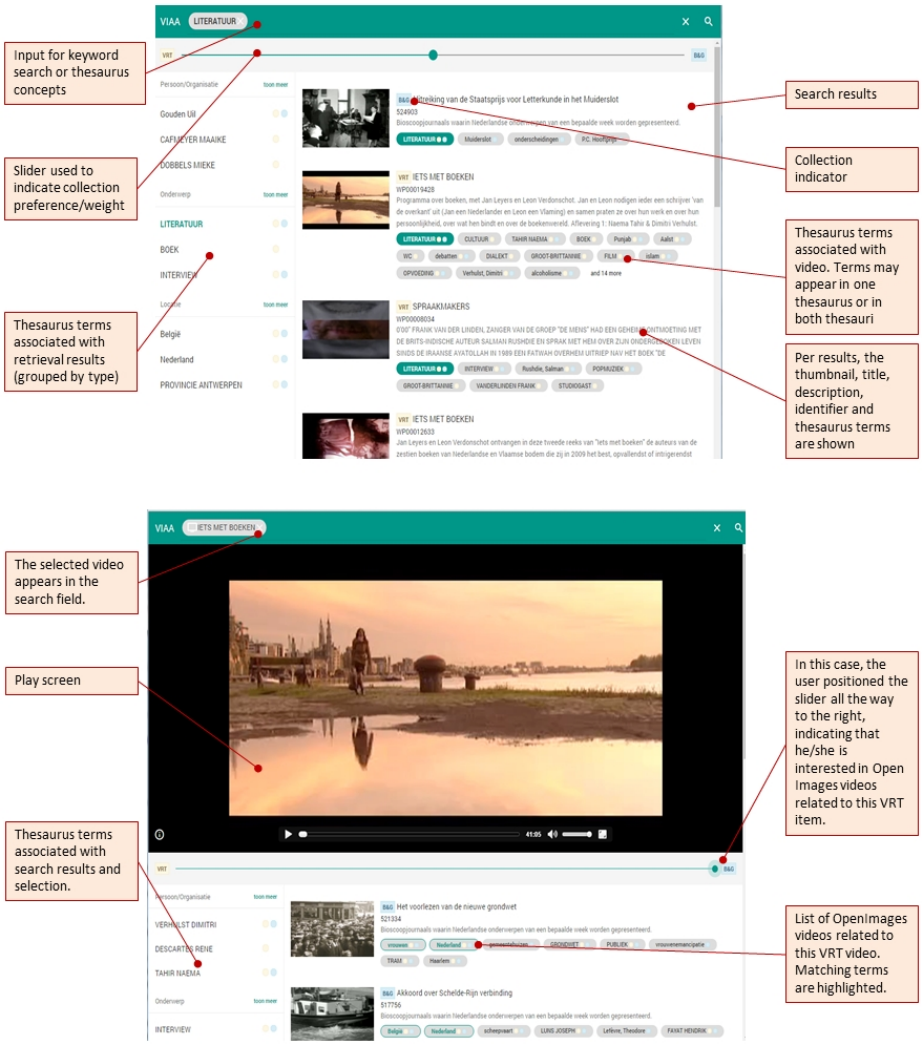

Fig. 2. Annotated screenshot of the demonstrator

selected, after which they are used as search terms. The top screenshot shows an example where the concept "Literature" is selected[17]. This concept is present in both thesauri and a link was established in the alignment (both a blue and yellow marker are shown). The demonstrator shows results from both collections. At the top of the screen a slider can be found, which allows users to adjust the weight given to one of the two collections. This is an important feature of the demonstrator, as it allows users to actively find relations between items *across* the collections, even when initially one collection might have many more "hits". When a user has selected a video (bottom screenshot), this video is presented to the user (full screen if preferred). Below the video, related videos are shown. These are determined through their corresponding thesaurus concepts.

4.2 Demonstrator backend

The demonstrator takes as input a query, which is any combination of keywords, selected thesaurus terms or selected items. The application then presents as search results the best matching videos of the two collections, based on their annotations with concepts from the two thesauri.

The application can be considered an information retrieval application (a search engine), where the desired results would be evaluated on their relatedness to the user query. The demonstrator is constructed using the "search by strategy" approach[5], in which the back end developer can connect visually a variety of search-related components to design a *search strategy* that specifies how to retrieve relevant videos[18]. Figure 3 shows the search strategy designed for the search and suggestion functionality for VIAA application. We explain the elements below.

Data Source The strategy starts top left with a 'Data Source' block, that represents the entire database. The corresponding data consists of the two thesauri, the two sub-collections, and the link sets resulting from the alignment. In the application, we only search for videos, in our case identified as instances of the class `http://schema.org/VideoObject`. The search strategy uses a filter to only retrieve objects of this type Video.

Inputs In the application, the user can create a complex query consisting of a combination of keywords, thesaurus concepts, as well as an example video. In the strategy these inputs are represented by the green-labeled blocks at the top.

Search The three types of inputs from the search request contribute to the search results in different ways. The keywords are used to search the titles and descriptions of the videos (four blocks on the left side of Figure 3). Also, the keywords are used to locate thesaurus concepts, which then lead through the subject relation to relevant videos. The concepts in a search query may lead directly to a video related through the subject relationships.

[17] Note that the thesaurus labels are in Dutch which we translate for this paper.

[18] Specifically, the application is implemented using the Spinque Core platform `http://www.spinque.com/spinque-core`

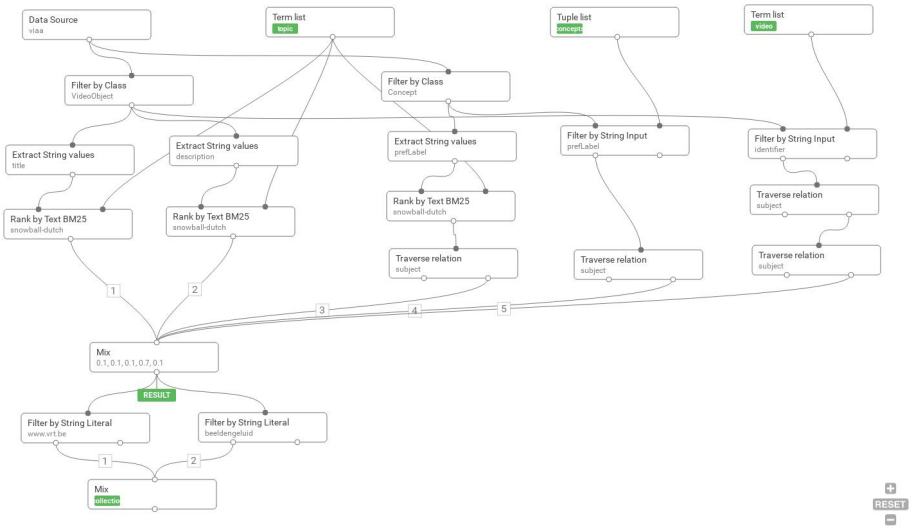

Fig. 3. Screenshot of the search strategy used in the backend of the demonstrator.

If a user-selected video is a part of the search query, then the goal is to find videos related to that video. In order to achieve this, the thesaurus concepts are used. This is done with the blocks on the right side of the strategy shown in Figure 3. Concepts with which the selected video is annotated are selected, and videos described with the same concepts are retrieved from the collection. The related videos are weighted by the number of concepts matches: the more overlap, the more relevant the results.

Combining results The sub-results of the various sub-strategies are combined in a "mix block". This block has 5 inputs, each with a weight that indicates how much should be included. In our application, results found through matching concepts are considered more important than the results using keywords and thus receive a higher weight.

Collection weighing The results contain videos of the VRT collection as well as the OpenImages collection. The last three blocks determine the priority of each collection in the results. In the search application, the user can specify (with a slider) which collection should be emphasized more for that specific search query. In the last mix block this user input is used to set the weights. Weights can be set between 0-1 with the total weight of the two collections adding up to 1.

5 Discussion

Converting the VRT thesaurus The conversion of the VRT thesaurus was done by IT specialists. However, knowledge of the internal structure of the source thesaurus provides significant time savings and quality assurance. Combining

"technical know-how" (about XML, RDF, SKOS) and "content know-how", we estimate that a conversion project such as this can be completed in about 5 working days. This indicates that it is a viable option for institutions to work together with specialists to convert their thesaurus. Using the produced scripts, any updates on the datamodel or actual content are easily carried through to the SKOS version. For completely new thesauri, the scripts can be used as starting points, saving time. The work performed may thus serve as a basis for publishing a linked thesaurus. The thesaurus could be published online as Linked Open Data using a specific solution, such as the OpenSKOS platform[19], which is used by the GTAA. The experience gained during this project shows us not only some possible further steps, but can also be used as a use case to convince other organizations to disclose current non-standardized and / or published thesauri as linked data.

Mapping strategies and generated links In general, aligning thesauri is still a daunting task, even when the labels of the terms are available in the same language (in our case Dutch). A variety of reasons may underly the way different structures are arranged in the thesauri[4]. In our specific case, it was non-trivial to identify the corresponding parts between two thesauri (Persons, Places etc.) as both thesauri took different modeling decisions and have a different structure. The fairly 'flat' GTAA is at the highest level divided in a number of concept schemes whereas the VRT thesaurus has a lot or hierarchical structure, but no concept schemes. However, by allowing the application of different filtering options and fine-tuning these interactively, CultuurLINK makes it possible to identify these corresponding parts. For each of the parts, specific sub-strategies could be made using different label matchers. Even though the thesauri have different origins, structure and usage, we still find large numbers of links between the two. In addition, it is likely that the four strategies which developed within this project serve as blueprints for mapping strategies between other thesauri.

Here too, a mixed team is in the best position to find the links. Ideally combining content managers and IT people, aware of the (dis)advantages and workings of strings matchers, structural matchers, fuzzy matching, etc. Even so, an interactive, user-friendly tool such as CultuurLINK allows for users with less technical knowledge to still produce good alignments.

The links produced should be further explored to determine and possibly expand the coverage and quality. This process is guided by revisiting unmapped concepts to determine whether these are terms that have no counterpart in the other thesaurus, or they should be classified as errors.

The links between the VRT thesaurus and GTAA act as a possible bridge to the larger Web of Data[3]. In previous projects, the GTAA thesaurus has been linked to other thesauri and datasets, including Wordnet[20] and DBpedia[21] (cf. [6], [7]). Through the alignment described here, the VRT thesaurus and collection

[19] `http://openskos.beeldengeluid.nl/`
[20] `https://wordnet.princeton.edu/`
[21] `http://dbpedia.org`

are now linked to these sources as well. Furthermore, these existing links can also be exploited to either add or verify the links between VRT thesaurus and GTAA.

In the current form, we only produce and show exact matches. However, we could also consider close matches, based on near-exact string matches for example. These could also be used in the demonstrator when no exact matches are found.

Extension and reusability of the demonstrator At this time, the demonstrator covers two subsets of the much larger Flemish and Dutch collections guarded by the institutions. Both VIAA and NISV work towards making the metadata of a much larger set of items available as Linked Data, using the thesauri described here. As soon as this metadata becomes available, it can be added to the dataset of the demonstrator to encourage wider usage and retrieval results.

6 Related Work

Prior descriptions of tools and use cases for linking cultural heritage thesauri exist. Van Assem et al. describe a method for converting thesauri to SKOS[8], while the museum use case described in [9] discusses a similar approach to conversion. Cross-collection browsers that exploit linked thesauri have been previously explored in the MultimediaN Eculture [10] and MuseoFinland[11], but with less emphasis on exploring links between two collections. The DIVE linked media browser[12] partially overlaps in terms of collections, also enabling browsing of the Open Images collection and the links to other collections. Here, also the user has limited control over the retrieval of related objects from two collections. The level of user control in our demonstrator allows for more effective exploration of cross-collection links.

7 Conclusions

The case study shows how we combine existing and new tools to provide integrated browsing and search across different (sub)collections from two national audiovisual archives. We describe the conversion of a legacy thesaurus and present the reusable conversion algorithms. We illustrate the benefits of interactive alignment using the CultuurLINK tool; specifically, we show how filters can be used to isolate different sections of each thesaurus, so that *section-specific* matchers can be employed. This alignment strategy uses a combination of filtering and matching techniques that work for these two specific thesauri and the exact same strategy will not work (as well) for two different thesauri. This is precisely why a transparent, interactive alignment method (and tool) is needed. The converted thesaurus and the links produced are represented as SKOS RDF files that can be accessed online for easy reuse.

We finally introduce a cross-collection browser which uses the produced alignment to allow users to explore connections between the two collections. This application uses a flexible search strategy to retrieve relevant items from the

two collections based on user search queries in the form of keywords, thesaurus terms or selected videos. The tool features a slider to allow users to put more emphasis on one or the other collection.

Acknowledgments This research was funded by Taalunie project "Gemeenschappelijke Thesaurus voor Uniforme Ontsluiting".

References

1. Bron, M., van Gorp, J., Nack, F., de Rijke, M., Vishneuski, A., de Leeuw, S.: A subjunctive exploratory search interface to support media studies researchers. In: Proceedings of SIGIR12. (2012)
2. Miles, A., Matthews, B., Wilson, M., Brickley, D.: Skos core: simple knowledge organisation for the web. In: International Conference on Dublin Core and Metadata Applications. (2005) pp–3
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
4. van Ossenbruggen, J., Hildebrand, M., de Boer, V.: Interactive vocabulary alignment. In: TPDL. (2011) 296–307
5. de Vries, A.P., Alink, W., Cornacchia, R.: Search by strategy. In: Proceedings of the third workshop on Exploiting semantic annotations in information retrieval, ACM (2010) 27–28
6. Bouma, G.: Cross-lingual ontology alignment using eurowordnet and wikipedia. Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010) (2010) 1023–1028
7. Malaisé, V., Isaac, A., Gazendam, L., Brugman, H.: Anchoring dutch cultural heritage thesauri to wordnet: two case studies. ACL 2007 (2007) 57
8. van Assem, M., Malaisé, V., Miles, A., Schreiber, G.: A method to convert thesauri to skos. In Sure, Y., Domingue, J., eds.: The Semantic Web: Research and Applications. Volume 4011 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2006) 95–109
9. de Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. In: Extended Semantic Web Conference, Springer (2012) 733–747
10. Schreiber, G., Amin, A., Van Assem, M., De Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., et al.: Multimedian e-culture demonstrator. In: International Semantic Web Conference, Springer (2006) 951–958
11. Hyvönen, E., Mäkelä, E., Kauppinen, T., Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., et al.: Culturesampo: A national publication system of cultural heritage on the semantic web 2.0. In: European Semantic Web Conference, Springer (2009) 851–856
12. de Boer, V., Oomen, J., Inel, O., Aroyo, L., Van Staveren, E., Helmich, W., De Beurs, D.: Dive into the event-based browsing of linked historical media. Web Semantics: Science, Services and Agents on the World Wide Web **35** (2015) 152–158