Undefined 0 (0) 1
IOS Press

Amsterdam Museum Linked Open Data

Victor de Boer [a,*] Jan Wielemaker [a] Judith van Gent [b] Marijke Oosterbroek [b] Michiel Hildebrand [a] Antoine Isaac [a,c] Jacco van Ossenbruggen [a] Guus Schreiber [a]

[a] *Department of Computer Science, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*
Email: {v.de.boer, j.wielemaker, m.hildebrand, j.r.van.ossenbruggen, guus.schreiber}@vu.nl
[b] *Amsterdam Museum, Postbus 3302, 1001 AC Amsterdam, The Netherlands*
Email: {j.vangent, m.oosterbroek}@amsterdammuseum.nl
[c] *Europeana, Koninklijke Bibliotheek, Prins Willem-Alexanderhof 5, 2509 LK, Den Haag, The Netherlands*
Email: antoine.isaac@kb.nl

Abstract. In this document we describe the Amsterdam Museum Linked Open Data set. The dataset is a five-star Linked Data representation and comprises the entire collection of the Amsterdam Museum consisting of more than 70,000 object descriptions. Furthermore, the institution's thesaurus and person authority files used in the object metadata are included in the Linked Data set. The data is mapped to the Europeana Data Model, utilizing Dublin Core, SKOS, RDA-group2 elements and the OAI-ORE model to represent the museum data. Vocabulary concepts are mapped to GeoNames and DBpedia. The two main contributions of this dataset are the inclusion of internal vocabularies and the fact that the complexity of the original dataset is retained.

Keywords: Cultural Heritage, Museum, Thesaurus, Europeana Data Model

1. Introduction

In this document, we describe the Amsterdam Museum Linked Data set. The Amsterdam Museum[1] is a Dutch museum hosting objects related to the history and culture of Amsterdam and its citizens. Among these objects are paintings, drawings, prints, glass and silver objects, furniture, books, costumes, etc. At any given moment, around 20% of the objects are on display in the museum's exhibition rooms, while the rest is stored in storage depots.

As do many museums, the Amsterdam Museum uses a digital data management system to manage their collection metadata and authority files, in this case the proprietary Adlib Museum software[2]. As part of the museum's policy of sharing knowledge, in 2010, the Amsterdam Museum made their entire collection available online using a creative commons license. The collection can be browsed through a web-interface[3]. Second, for machine consumption, an XML REST API was provided that can be used to harvest the entire collection's metadata or retrieve specific results based on search-terms such as on creator or year[4]. The latter API has been used extensively in multiple Cultural Heritage-related application-building challenges[5].

While larger cultural heritage institutions such as the German National Library[6] or British National Library[7] have the resources to produce their own Linked Data, smaller institutions often depend on large-scale aggregators such as Europeana. Europeana aggregates metadata from more than 2,200 European cultural heritage institutions and provides access through its

*Corresponding Author

[1] http://amsterdammuseum.nl
[2] http://www.adlibsoft.com/
[3] http://collectie.amsterdammuseum.nl/
[4] An example is http://amdata.adlibsoft.com/wwwopac.ashx?database=AMcollect&search=creator=Helst*
[5] http://blog.amsterdammuseum.nl/?p=5245 (Dutch)
[6] https://wiki.dnb.de/display/LDS/
[7] http://bnb.data.bl.uk

Web portal[8]. The Europeana Linked Data pilot [9] uses metadata from 200 of these institutions and provides Linked Open Data access, which is described in the data.europeana.eu paper in this special issue [7]. In current workflow of this pilot, metadata records are ingested and restructured to fit the Europeana Data Model (EDM) [5] and published it on Europeana servers as "five-star" Linked Data [1]. Although this approach ensures a level of consistency and interoperability between the datasets from different institutions the restructuring creates a disconnect between the cultural heritage institute original metadata model and the Linked Data version.

A large part of the research that has resulted in the dataset described in this document was carried out within the context of EuropeanaConnect[10]. Within EuropeanaConnect, different technologies and components for Europeana were developed, including the methodology and tools of which the Amsterdam Museum Linked Data set is the result. The dataset has been included in the Europeana Thoughtlab, a set of innovative technologies and tools that lead the way for Europeana-related developments[11].

The Amsterdam Museum Linked Data set was created with a number of design principles in mind. Specifically, we have created a five-star Linked Data set for a small museum that 1) retains the complexity of the original data while 2) achieves interoperability through a mapping to an interoperability layer - in this case the EDM. The Amsterdam Museum Linked Data set implements best practices that, together with its methodology and tools, Europeana is keen on adopting for its future workflow. This is exemplified by the inclusion of the dataset in the Europeana Thoughtlab and by the adoption of specific modeling choices in future versions of the Europeana Data Model (for example, the method of achieving interoperability as described in Section 2.3 is adopted in [5, Sect. 5.4]).

2. Metadata Conversion and Modeling

2.1. Conversion

We here describe briefly the process used to convert the original data to Linked Data. The methodology and tools that focus on a high level of interactivity and transparency of the process are described in more detail in [2].

The Amsterdam Museum data consists of three parts: 1) an object metadata set consisting of metadata records for the 73,447 objects; 2) a thesaurus consisting of 28,000 concepts used in the metadata records and 3) a person authority file consisting of 66,966 persons related to the objects or the metadata. The metadata, thesaurus and person authority file were all harvested through an OAI-PMH interface [12]. The resulting XML was first converted to crude RDF and subsequently restructured using interactive rewriting rules. This was done with the XMLRDF tool [2]. Resources (Objects, concepts, persons, ...) were assigned URIs. URIs are made up of three parts. For the URI basename, we used `http://purl.org/collections/nl/am/`, here shortened to `am:`. Second, a term denoting the type of resource, followed by a dash: 'proxy-', 'aggregation-', 't-' or 'p-' for proxies, aggregations, concepts and persons respectively. Finally, the URI ends with a unique Amsterdam Museum identifier (the 'priref'). An example URI is `http://purl.org/collections/nl/am/proxy-22476` which has `am:proxy-22476` as a shorthand notation. For predicate URIs, we use the basename plus the original XML element names (eg. `am:maker`). At the time of conversion, we did not have access to the Amsterdam Museum web servers. This meant we could not have URIs in the Amsterdam Museum namespace redirect to our semantic server so that they can be resolved. We therefore used purl.org URIs.

Also in this step, implicit links (for example between between objects and thesaurus concepts) are made explicit as RDF relations. Literal values that represent references to resources are replaced by those resources. Language information of the literals is also added to the data in the form of language-typed literals. The internal links are described in Section 3.4. Figure 1 shows examples of both the internal links and the language-typed literals.

2.2. The Europeana Data Model

Within the cultural heritage domain, a number of metadata schemas exist. Popular schemas for museums are models such as Dublin Core (DC) [6], Visual

[8] `http://www.europeana.eu`
[9] `http://data.europeana.eu`
[10] `http://www.europeanaconnect.eu/`
[11] `http://pro.europeana.eu/web/guest/thoughtlab/linked-open-data`

[12] Open Archives Initiative - Protocol for Metadata Harvesting: `http://www.openarchives.org/pmh/`

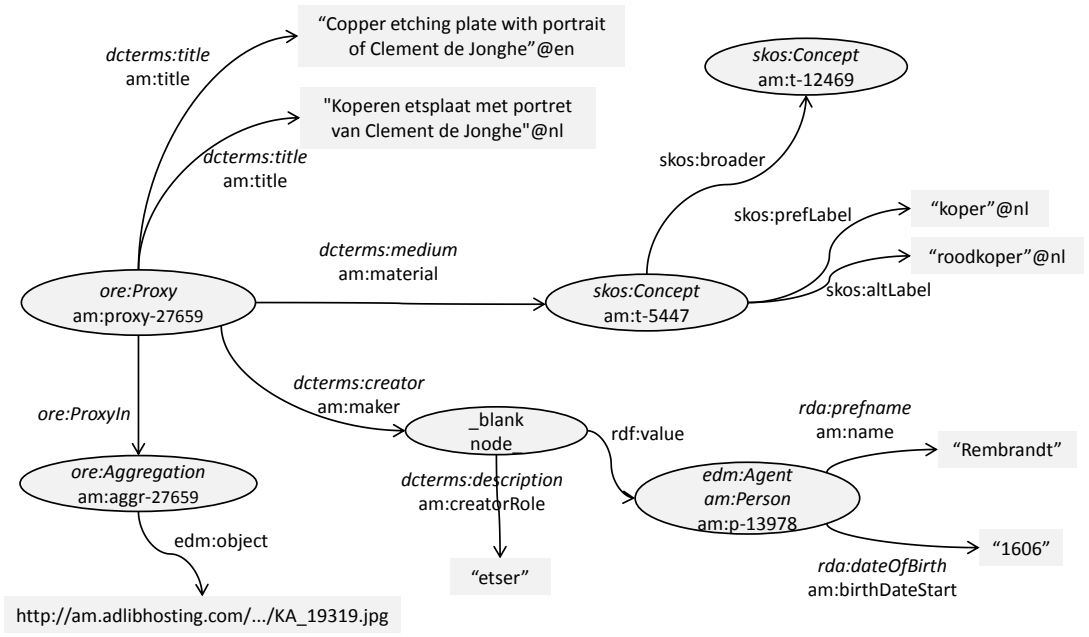

Fig. 1. A small part of the RDF graph surrounding the resource `am:proxy-27659`. Predicates and classes are listed with their super-properties and super-classes in italics. On the left side, the relation to the Aggregation (with the attached thumbnail) can be seen. In the top right, the relation to the thesaurus concept is shown. In the bottom right, a complex creator relation (the relation has a type as well as a value) requires a blank node. The `rdf:value` of the blank node is a resource in the Person list. The proxy has two title triples, for the English and Dutch titles respectively, with language-typed literals as objects.

Resources association (VRA)[13], the Lightweight Information Describing Objects schema (LIDO) [9] or the CIDOC Conceptual Reference Model (CRM) [3]. EDM is not built on any particular community standard but rather adopts an open, cross-domain Semantic Web-based framework that can accommodate the range and richness of these schemas. It has been tested for compatibility with other community standards such as the Encoded Archival Description (EAD)[14] for archives and the Metadata Encoding and Transmission Standard (METS)[15] for digital libraries [4].

In fact, EDM mainly re-uses or draws inspiration from elements belonging to other standards. DC, CIDOC-CRM, SKOS are used for "descriptive" metadata.[16] For person metadata, EDM also uses the Resource Description and Access (RDA) Group 2 metadata standard[17]. These properties include given and family names, birth and death dates etc.[18] For more "technical" and "organization-related" metadata aspects, requirements specific to large-scale aggregation and access to digitized resources have been taken into account, making EDM a fairly unique proposal as these scenarios emerged only recently. EDM for example supports multiple providers describing the same object and allows for enrichment of the museum data, while clearly showing the provenance of all the data that links to digital objects. This is achieved by incorporating the *proxy-aggregation* pattern from the Object Re-use and Exchange (ORE) model [8]. For more details and discussion on EDM we refer the reader to the paper on the data.europeana.eu dataset in this special issue [7].

[13] `http://www.vraweb.org/projects/vracore4/`
[14] `http://www.loc.gov/ead/`
[15] `http://www.loc.gov/mets/`
[16] To a great extent, our interoperability approach is therefore easily transferable to application contexts that exploit other descriptive metadata element sets.
[17] `http://rdvocab.info/ElementsGr2`
[18] Europeana's choice for RDA was partly informed by the research presented in this paper.

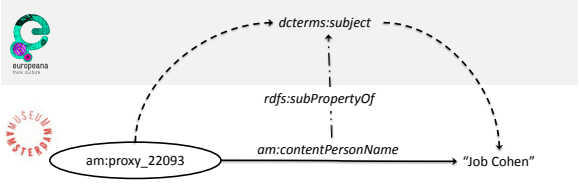

Fig. 2. RDF fragment showing how metadata mapping ensures interoperability.

2.3. Mapping to the Europeana Data Model

To make the Amsterdam Museum Linked Data interoperable with the EDM, two steps are taken. First of all, the museum objects are represented as proxy-aggregation pairs. For our purpose, this means that an Amsterdam Museum metadata record gives rise to both a proxy resource as well as an aggregation resource. The RDF triples that make up the object metadata (creator, dimensions etc.) have the proxy as their source while the triples that are used for provenance (data provider, rights etc.) as well as digital representation (e.g. thumbnails) have the aggregation resource as their source. An example of this is shown in Figure 1. The bottom part of the figure shows an example triple relating an object to the name of a depicted person. DC only has a single notion of the subject of a work. By mapping the specific properties to the more general property using the `rdfs:subProperty` in the metadata schema, an application capable of RDFS reasoning can infer that the object has "Job Cohen" as its subject. We therefore achieve interoperability without discarding the complexity of the original data.

The second step is to map the Amsterdam Museum specific classes and properties to those of the EDM using `rdfs:subPropertyOf` or `rdfs:subClassOf` relations through a schema file. Specifically, we map the Amsterdam Museum properties and classes to those from DC, SKOS, RDA and to EDM-specific elements. Through these mappings, interoperability of the museum-specific metadata with the EDM is achieved. An example is shown in Figure 2.

3. Description of the Linked Data set

We now describe in more detail the resulting Linked Data set, for each of its three parts. Table 1 lists some statistics. For illustration, Figure 1 shows a small part of the RDF graph for a museum object, including internal links (relations to Amsterdam Museum resources).

Data part	*Resources*	*Predicates used*	*RDF triples*
Object metadata	73,447 (proxies)	100	5,700,371
Thesaurus	28,000 (concepts)	13	601,819
Person auth. list	66.966 (persons)	21	301,143
Total	168,413	134	6,603,333

Table 1
Some statistics for the three parts of the Linked Data set

3.1. Object Metadata

The object metadata consist of 73,447 proxy-aggregation pairs. The 100 different predicates include creator, dimensions, locations, related exhibitions etc. For complex relations, 566,239 blank nodes were retained. In total the object metadata consists of 5,700,371 RDF triples. 975,859 triples have a thesaurus concept as object and 210,407 triples have a person resource as object.

The RDFS mapping file relates the 100 Amsterdam Museum properties to the EDM properties through the `rdfs:subPropertyOf` construct. 90 properties are defined as subproperties of DC properties, seven properties are mapped to EDM-specific properties (`edm:hasMet` , `edm:happenedAt`, etc.) and three properties are defined as subproperties of `rdfs:label`. Two Amsterdam Museum classes `am:Exhibition` and `am:Locat` were defined as `rdfs:subClassOf` of the EDM class `edm:Event`. Instances of these two classes are used to describe Amsterdam Museum-specific events related to the cultural heritage object.

3.2. Thesaurus

The thesaurus consists of 28,000 concepts represented in SKOS. These include geographical terms, motifs, events etc. Most term-based thesauri, including the AM thesaurus, have a more or less uniform structure, based on the ISO 2788 standard[19]. This standard defines two types of terms (preferred and non-preferred) and five relations between terms: broader, narrower, related, use and use for. Use and use for are allowed between preferred and non-preferred terms, the others only between preferred terms. Translating thesauri from this format to SKOS is fairly straightforward and well documented (cf. [10]). There are 13 predicates including `skos:broader` (7,487 triples) and `skos:narrower` (8,486 triples) that establish hier-

[19] `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=7776`

archies. In total the thesaurus consists of 601,819 RDF triples (no blank nodes).

3.3. Person authority file

The person authority file consists of 66.966 instances of `am:Person` (a subclass of `edm:Agent`). The persons in this data set are creators, past or present owners of art objects, annotators, depicted persons etc. In this case, the original 21 distinct Amsterdam Museum predicates were used. These predicates include birth and death dates, nationality, alternative name spellings etc. These properties are mapped to RDA Group 2 elements using 20 `rdfs:subProperty relations`. The `am:Person` class was also mapped as a `rdfs:subClassOf edm:Agent`. In total there are 301,143 RDF triples in Person data set.

3.4. Internal Links

There are many links between resources in the dataset. There are 558,161 links between the 73,447 proxies to thesaurus concepts, producing an average of 7.6 links to thesaurus concepts per cultural heritage object. Examples of such links are `am:material` and `am:contentMotifGeneral`. There are also 80,432 links between proxies and persons, mostly through the `am:contentPersonName` and `am:associationPersonName` properties. In addition to this, there are 243,532 links between proxies and other proxies, using for example the `am:relatedObjectReference` property.

3.5. External Links

Links to external data sources were established manually using the Amalgame alignment platform [11]. Amalgame supports an interactive alignment process where the user iteratively applies can apply different (string) matching techniques, mapping partitioners and filters. After each step the user typically analyzes the results and can adjust settings or prepare next steps. Matchers, filters etc. are combined into an alignment workflow that can be run on the vocabularies that are to be linked. The Amsterdam Museum alignment process is described in more detail in [2], we here give a high-level description.

For the Amsterdam Museum thesaurus, the final workflow includes a type-based filter that splits the concepts into geographical and non-geographical concepts. The non-geographical concepts were mapped to the Dutch Art and Architecture thesaurus (AATNed)[20] using a simple exact string matching algorithm. Of the resulting mappings, the non-ambiguous mappings are separated from the ambiguous mappings (where one Amsterdam Museum concepts is mapped to multiple target concepts). A manual evaluation of a small subset indicated that the precision is around 80%. These non-ambigous matches (2,498 links) were added to the data.

For the geographical concepts, a similar workflow was employed to map the concepts to GeoNames[21]. An evaluation of a subset of the produced links showed that the precision is above 90%. The 143 covering about 75% of the geographic part of the Amsterdam Museum thesaurus.

The person authority file was mapped to Getty Union List of Artist Names (ULAN)[22]. Here, the Amalgame workflow included first matching based on canonical person names and secondly matching on alternate spellings of the name. Resulting mappings were also split based on ambiguity. Through evaluation of subsets, we identified three categories of mappings: high precision (100% correct), mid-precision (80-90% correct) and low precision (<80%). The high- and mid-level mappings to ULAN were added to the data resulting in 1,426 links. Finally 34 persons were linked to persons in DBpedia. This is a relatively low number as 1) most of the Amsterdam Museum people are not well-known enough to appear in DBPedia and 2) we used fairly simplistic matching algorithms here.

In total 3,753 links to external data sources are included. Although this is only a fraction of the total number of concepts, the usefulness of these mappings is much greater as they represent the part of the concepts with which the most metadata is annotated. In total, 70,742 out of the 73,447 (96%) objects are annotated with one or more concepts or persons that have been linked, with an average of 4.3 linked concepts per object. Nevertheless, we are still aiming to enrich the data with more links.

4. Availability

The Amsterdam museum data, consisting of the converted datasets, the schema mapping files and the

[20] `http://www.aat-ned.nl`
[21] `http://www.Geonames.org`
[22] `http://www.getty.edu/research/tools/vocabularies/ulan`

high-quality mapping files are served as Linked Open Data on the Europeana Semantic Layer (ESL). The ESL is a running instance of the ClioPatria semantic server [12] that houses other datasets that have been mapped to EDM. It can be accessed at http://semanticweb.cs.vu.nl/europeana. Amsterdam Museum PURL URIs (for example http://purl.org/collections/nl/am/proxy-63432) are redirected to this server. Based on the response header in the HTTP request, either HTML, RDF/XML or RDF/Turtle is served. The SPARQL endpoint for the ESL is found at http://semanticweb.cs.vu.nl/europeana/sparql/[23]. In addition to this, the Amsterdam Museum dataset can be separately accessed through a GIT repository at http://eculture.cs.vu.nl/git/public/?p=econnect/metadata/AHM.git. More information, including example URIs and SPARQL queries can be found at http://semanticweb.cs.vu.nl/lod/am.

5. Discussion

The Amsterdam Museum Linked Data set is a significant data source for Amsterdam history and culture. In previous application development competitions such as Apps for Amsterdam[24], the Amsterdam Museum dataset has been used extensively for a number of mashup applications and we expect that the Linked Data version will be an equally central data set for the web of cultural heritage Linked Data.

Future work on the data set includes efforts to produce more links, both to Amsterdam and Dutch cultural heritage datasets and vocabularies as well as to more general vocabularies such as VIAF or DBPedia. At the same time, we plan to validate our design choices by developing a number of Web and mobile applications that combine the Amsterdam Museum data with other datasets. One example is a mobile cultural tour guide for the city of Amsterdam in which the Amsterdam Museum Linked Data set will be a central data source, which we are currently developing. In another project, we combine the Amsterdam Museum data with World War II archival and library data. We expect that these efforts will also provide us with feedback on the specific design choices made, which can inform next version.

The Amsterdam Museum Data Set was developed in the context of the EuropeanaConnect research project. It serves as a prototype Linked Data set in which the original richness of the data is maintained, while still being interoperable through its mapping to the EDM. A limitation for direct re-use in Europeana is that this interoperability does require some RDFS reasoning by applications wishing to access the data at the interoperability level.

Another difference with the current Linked Data pilot of Europeana is that the latter focuses on producing a Linked Data set based on the already-ingested metadata, This metadata consists of a minimal set of Dublin Core properties. Current research within Europeana is gearing towards harvesting and producing more complex data sets, in the manner of the Amsterdam Museum one. When more of these data sets will appear, more links to Amsterdam Museum data are likely.

In the current workflow, the Linked Data is not updated automatically once a change has been made in the original collection management system. The OAI-PMH harvesting, conversion and data uploading steps would have to be run again. The XMLRDF conversion rules as well as the Amalgame alignment workflows are are retained and can be re-run on new data at any moment. The performance is linear in the amount of input records. On a dual CPU P8700 @ 2.53 GHz with 4 GB of RAM, the initial conversion to crude RDF of 73,447 records takes 117.02 seconds. Executing the 58 XMLRDF rewriting rules on this crude RDF takes 89.78 seconds. The total conversion time for this dataset is 206.8 seconds.

For Amsterdam Museum, this conversion can be executed in an automated fashion when the original data is updated. The conversion tool ensures that the same resources receive the same URIs. The workflow has partial support for incremental updates. When new records are added, they can be processed incrementally by the conversion and linking tools to produce new RDF triples. However, currently the XMLRDF tool does not implement functionalities to adopt modified or deleted records incrementally. In this case, it is necessary to run the conversion script on the entire dataset.

Currently, the Amsterdam Linked Data set is hosted on the ESL. We are looking at the feasibility of having the data hosted by the Amsterdam Museum itself, which could contribute to persistency and maintainability of the data. The use of PURL URIs allows us to redirect HTTP requests to another server, when this happens.

[23] An interactive SPARQL query page is available at http://semanticweb.cs.vu.nl/europeana/user/query

[24] http://www.appsforamsterdam.nl/

References

[1] Tim Berners-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/ LinkedData.html, 2006.

[2] Viktor de Boer, Jan Wielemaker, Judith van Gent, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. Supporting linked data production for cultural heritage institutes: The Amsterdam Museum case study. In Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti, editors, *ESWC*, volume 7295 of *Lecture Notes in Computer Science*, pages 733–747. Springer Berlin / Heidelberg, 2012.

[3] Martin Doerr. The CIDOC CRM - an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24:2003, 2003.

[4] Martin Doerr, Stefan Gradman, Steffen Hennicke, Antoine Isaac, Carlo Meghini, and Herbert van de Sompel. The europeana data model. Dissemination paper. IFLA 2010. World Library and Information Congress: 76th IFLA General Conference and Assembly. Gothenburg, Denmark, 15 August 2010. http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf, 2010.

[5] Antoine Isaac (ed.). Europeana data model primer. Europeana technical document http://pro.europeana.eu/edm-documentation, 2010.

[6] Dublin Core Metadata Initiative. Dublin core metadata element set version 1.1. http://dublincore.org/documents/1999/07/02/dces/, 1999.

[7] Antoine Isaac and Bernhard Haslhofer. Europeana linked open data - data.europeana.eu. *Forthcoming in Semantic Web Interoperability, Usability, Applicability http://www.semantic-web-journal.net/content/europeana-linked-open-data-%E2%80%93-dataeuropeanaeu*, 2012.

[8] Carl Lagoze, Herbert Van de Sompel, Michael L. Nelson, Simeon Warner, Robert Sanderson, and Pete Johnston. Object re-use & exchange: A resource-centric approach. Arxiv preprint. arXiv:0804.2273v1 ; http://arxiv.org/abs/0804.2273, 2008.

[9] Richard Light, Gordon McKenna, Regine Stein, and Axel Vitzthum. LIDO - lightweight information describing objects. ATHENA project deliverable D3.3, http://www.athenaeurope.org/getFile.php?id=535, 2009.

[10] Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber. A method to convert thesauri to SKOS. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 95–109. Springer Berlin / Heidelberg, 2006.

[11] Jacco van Ossenbruggen, Michiel Hildebrand, and Viktor de Boer. Interactive vocabulary alignment. In Stefan Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt, editors, *Research and Advanced Technology for Digital Libraries - International Conference on Theory and Practice of Digital Libraries, TPDL 2011, Berlin, Germany, September 26-28, 2011. Proceedings*, volume 6966 of *Lecture Notes in Computer Science*, pages 296–307. Springer Berlin / Heidelberg, 2011.

[12] Jan Wielemaker, Michiel Hildebrand, Jacco Ossenbruggen, and Guus Schreiber. Thesaurus-based search in large heterogeneous collections. In A.P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, editors, *The Semantic Web - ISWC 2008, Proceedings of the Seventh International Semantic Web Conference*, volume 5318 of *Lecture Notes in Computer Science*, pages 695–708, Berlin, Heidelberg, 2008. Springer Berlin / Heidelberg.