Towards the Development of a Metadata Model for a Digital Cultural Heritage Collection with Focus on Provenance Information

Susanne Al-Eryani
State and University Library Göttingen
salerya@sub.uni-goettingen.de

Stefanie Rühle
State and University Library Göttingen
sruehle@sub.uni-goettingen.de

Abstract

This project report describes the first steps of the development of a metadata model for the contextualization of heterogeneous objects from different cultural heritage collections with focus on provenance information. The project started with the assumption that aims and objectives of researchers working with cultural heritage collections differ from discipline to discipline. Accordingly, use cases and requirements for the description of objects are heterogeneous. To provide a model that would be usable not only within but also across academic disciplines the project needed to know where these requirements differ and where they match. Therefore the first part of the project was focused on the investigation of use cases and requirements. On the base of the common requirements a generic model will be build that allows the merging of data from a variety of disciplines using different metadata standards. The model's structure will be a combination of prevalent metadata standards mapped to each other. Another peculiarity of the model will be the modular design of micro-ontologies, sets of domain-specific class structures that are, nevertheless, available on a meta-level in terms of substructures. Applying the DCMI dumb-down principle these subproperties and subclasses will be assigned to a who-what-where-when model, a base structure for the description of objects.

The project divided the work process of the project into seven steps. As the project is still work in progress, only four steps will be explained in detail in this report. The three remaining steps will be presented in an outlook.

Keywords: metadata model; metadata standard; digital cultural heritage collection; provenance information

1. Introduction

The predicted shift or extension from the concepts of the current World Wide Web, the so called Web 2.0, to those of Web 3.0 is in full swing and the debates on how to establish an appropriate base in the context of this challenge often end up with a big question mark. The expression 'Web 2.0', coined by Darcy DiNucci (1999) and made popular by Tim O'Reilly (2005) at the Web 2.0 Conference, held in San Francisco in 2004, stands for an interactive medium that can be described as a web of documents connected by hyperlinks suitable for human consumption. In contrast, Web 3.0, also known as 'Semantic Web', is a web of structured data conveying semantic meaning and connected by semantically meaningful links. This web of machine-readable data will support people's needs, for example, to create data stores on the Web or to achieve precise information from an unmanageable number of options (see W3C, 2015). But, is there a golden road that would lead to a satisfying supply of information in the net in order to make data accessible and searchable according to most diverse requirements, and in addition, that would provide a base for embedding data appropriately into a semantic net of information? Or must metadata specialists and information professionals working in cultural heritage institutions develop their very special metadata model for "their own data" within "their own institution", when preparing the data for future requirements? However, the application of different standards leads to enormous challenges when it comes to the interlinking and automatic processing of data. The overall aim should be an extraction of the main concepts from the wide range of knowledge fields, transferred into a modest quantity of metadata schemes that would be sustainable and usable within different professional contexts. In a project, the work of which will be presented in this paper, the attempt is being made to create a metadata model that would meet this requirement.

The three-year project with its somewhat cumbersome designation Developing interoperable metadata standards for contextualizing heterogeneous objects, exemplified by objects of the provenance von Asch (short ASCH; see http://asch.wiki.gwdg.de) at the State and University Library of Göttingen (SUB), Germany, is lead by the SUB and the Institute of Social and Cultural Anthropology, in collaboration with the Metadata Group and the department Digital Library of the SUB and several collections of the Göttingen University (named in section 2), and is supported by the Deutsche Forschungsgemeinschaft (DFG). As the name implies, the project's work focuses on the development of a metadata model, which means the integration of various interoperable metadata standards (i.e. metadata schemes or element sets and corresponding application profiles) for the contextualization of heterogeneous objects of cultural heritage collections.

By describing resources of digital collections, the use of metadata standards and authority data is an essential technical precondition for making them identifiable and retrievable in the net. Considering the existing variety of information and the diverse but mostly not semantically defined web of relationships that link information to other information, the creation of a universally valid data model would be desirable but seems to be unconceivable, at least until present. Nevertheless, why should it not be possible to develop a very generic model for single segments of shared cultural knowledge that would be extensible in accordance to the requirements of individual research disciplines established in the various cultural heritage and scientific institutions? Would it be a realistic aim to generate a basic template that would be reusable and extendable in different contexts? Tim Berners-Lee formulates four assumptions for the interconnectedness of data: (1) things have to be named by URIs (Uniform Resource Identifier), (2) the URIs should be HTTP URIs, (3) useful information should be given on these URIs by using certain standards (RDF[1], SPARQL[2]), and (4) links to other URIs should be included (Berners-Lee, 2006). This linkable data, known as Linked Data, can be understood as a "set of best practices for publishing and connecting structured data on the Web using international standards of the World Wide Web Consortium" (Wood et al., 2014).

In order to provide linkable data, the resource descriptions must correlate to common metadata standards. A fast increasing number of scientific institutions, archives, libraries and museums are eager to prepare and edit their databases in order to make their digital resources interoperable even across institutional borders. This challenging task presupposes the application of appropriate metadata standards. There are a number of standards that fulfill the requirements of the different scientific and cultural heritage institutions: libraries are using, for example, MARC 21 and MODS (Metadata Object Description Schema), the application of EAD (Encoded Archival Description) and EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families) is commonly used by archives, and LIDO (Lightweight Information Describing Objects) is a widespread scheme applied by museums. In the world of natural science the ABCD (Access to Biological Collection Databases) and its extension ABCDEFG (ABCD Extended for Geosciences) are a first step to provide data across institutions and disciplines as is the Darwin Core standard. In addition to the application of metadata schemes, the use of authority data becomes indispensable for the description of resources because the semantic assignment via URIs facilitates an unambiguously identification of applied terms. Examples are the LCSH (Library of Congress Subject Headings), VIAF (Virtual International Authority File) and GND (Gemeinsame

[1] http://www.w3.org/RDF/

[2] https://www.w3.org/TR/rdf-sparql-query/

Normdatei)[3] used by libraries, the Getty Vocabularies[4] providing structured terminology for different cultural fields, and a wide array of taxonomies used in natural science.

The illustration of provenance information will be of special interest by developing the ASCH model. Therefore, common standards describing aspects of provenance will be considered. A variety of ways can be found for this description of resources, because different subject areas focus different aspects by documenting the life history of objects. T-PRO (Thesaurus der Provenienzbegriffe),[5] for example, is a thesaurus to describe terms of provenance in an object-orientated manner and is used by German libraries. For an event-based description of objects, LIDO is an appropriate format mainly used by museums. CIDOC CRM (CIDOC Conceptual Reference Model) enables to illustrate provenance information with the additional option of embedding evidences to the given facts, and an abstract level for description is possible by applying the PROV-DM (PROV Data Model)[6] provided by W3C.

The ASCH model is expected to merge different metadata standards commonly used by various cultural heritage institutions on a meta-level in order to make the metadata reusable in an interdisciplinary context as done by the DDB (Deutsche Digitale Bibliothek)[7] and Europeana,[8] for example. The who-what-where-when model developed by the DDB, and the Europeana Data Model (EDM) developed by Europeana allow specific object- and event-orientated resource descriptions, but provenance information cannot be illustrated in greater depth and an explicit interlinking to external evidence is not possible. To bridge the gap between these description frames is the aim of the ASCH project. The functionality of the ASCH model will be tested by using descriptions of digitized objects compiled from certain collections that are relevant for a chosen specific provenance context. The historical background of these collections will be depicted in the following section. Afterwards the methodology of the project's work will be explained in more detail.

2. Historical Background of the Collections

Seven collections of the Göttingen University are known to house or at least to have housed objects that were sent from Saint Petersburg in the second half of the eighteenth and the beginning of the nineteenth centuries. These collections are:

- the Historic Printed Collections, Manuscripts and Rare Books at the Göttingen State and University Library;
- the Ethnographic Collection at the Institute of Social and Cultural Anthropology;
- the Skull Collection at the Department of Anatomy and Embryology, Centre for Anatomy, University Medical Centre Göttingen;
- the Historical Collections at the Geoscience Centre;
- the Coin Cabinet at the Department of Archaeology;
- the Art Collection at the Department of Art History; and
- the Museum of Zoology.

The objects of these collections share a uniting circumstance in their life history because their provenance can be traced to a certain collector who had given them to a certain institution during a certain period of time. But, the characteristics of these objects are very distinctive and therefore they became prime candidates for the development of our metadata model. The objects' history leads us to the collector Georg Thomas von Asch (1729-1807), a Russian physician who had

[3] http://www.dnb.de/gnd
[4] http://www.getty.edu/research/tools/vocabularies/
[5] http://provenienz.gbv.de/T-PRO_Thesaurus_der_Provenienzbegriffe
[6] https://www.w3.org/TR/prov-dm/
[7] https://www.deutsche-digitale-bibliothek.de/
[8] http://pro.europeana.eu/

conducted his medical studies in Germany and had received his Doctorate of Medicine at the Georg August University in Göttingen. After his return to Russia, Baron von Asch had kept up close ties to Christian Gottlob Heyne (1729-1812), the director of the Göttingen University Library. The baron was also well acquainted with Johann Friedrich Blumenbach (1752-1840), the director of the Royal Academic Museum. Between 1771 and 1806, von Asch had sent more than 120 parcels and boxes to Göttingen, filled with natural and man-made objects of a wide range in order to be incorporated into the holdings of the University Library or the Academic Museum, respectively. In the second half of the nineteenth century, the Royal Academic Museum was dissolved and its collections were distributed among the new founded departments of the Göttingen University, named above, where they are partly to be found until present (for further readings see Hauser-Schäublin; Krüger [eds.], 2007).

In many cases, the origin of the ethnographica, botanica, zoologica, coins, rocks and other natural objects, skulls, prints, manuscripts, maps and books can be traced with great accuracy. Contemporary inventory books in the collection's archives, letters, especially the correspondence between Baron von Asch and Heyne, inventory lists enclosed to the parcels and boxes, and additional object descriptions, sometimes written on wrapping paper added by the donator, shed some light on the objects' biographies. In some cases, the provenance information is incomplete, e.g. object labels went lost during a flood, and some objects were given away in exchange for other objects so that their belonging to the former interdisciplinary collection cannot be proved. In other cases, provenance information on objects is available, but the current location of the items is unknown. Via preserved evidence it might be possible to reconstruct the objects' "journeys" and to bring them back virtually to their "home collections".

3. Work Methodology

Cultural and scientific heritage institutions have found their special way to manage collections by storing objects as well as information about objects. Analog formats for recording such as inventories, card catalogues, handwritten lists, vertical files and file labels can be found in the institutions' archives, but even in front of the doors of those buildings sometimes referred to as being old fashioned and dusty the technical revolution has not stopped. Meanwhile, analog recordings mostly have been transferred into a digitized format and stored objects have been photographed and digitized. Nowadays, the digitized metadata can increasingly be found in digital information systems that allow users access either open or locally restricted (Gilliland, 2008).

Turning to our purpose, what would be best practice to develop a single metadata model encompassing data received from different institutions with different research fields that handle their resource descriptions in various ways? In which manner could provenance information as well as external evidence referring to collection objects be linked? Our work methodology to achieve a solution can be reflected in the following seven steps which will be explained below:

1. empirical survey, analysis and evaluation of gathered information;
2. formulating of use cases;
3. analysis of requirements;
4. identification of classes and relations between classes;
5. identification of properties;
6. development of application profiles; and
7. testing the model's functionality.

3.1. Step One: Empirical Survey, Analysis and Evaluation of Gathered Information

Although we were eager to reuse widespread metadata standards and not to reinvent the wheel, we abandoned applying the complete element sets provided by these standards. Instead, we carried out an empirical survey in order to take the needs and requirements of various scholarly communities into account. Therefore, we conducted a two-day international workshop for which we invited about forty experts representing different scholarly disciplines (Anatomy, Archaeology, Computer Science, Geology, Geosciences, History, History of Art, Librarianship, Medicine, Mineralogy, Musicology, Social and Cultural Anthropology, Philology, and Zoology) who are known to be engaged in the subject matter of provenance. In small but heterogeneous focus groups as well as in the plenum we discussed questions and scenarios covering the following subjects: understandings of the term 'provenance', experiences with data bases and data exchange, use of authority data, handling of evidence proving an object's circle of live, practices concerning gathering and recording provenance information, reusability and editing of data, use of metadata standards and research infrastructures, and best practices, bad experiences and visions concerning work routines and research conditions. This form of information collection enabled us to obtain expert knowledge from representatives of the natural science and the humanities, and from archives, libraries, and museums as the three different kinds of cultural and scientific heritage institutions at once. An elaborated documentation of performance and topics of the workshop and detailed information on the round table discussions and the result analysis is available via the project's wiki (see http://asch.wiki.gwdg.de/index.php/Workshop_2015).

In addition to this workshop, about twenty one-on-one interviews with colleagues from institutions involved in provenance research or metadata creation gave us a greater understanding of their experience and wishes concerning provenance description.

Parallel to the empirical survey we analyzed the data provided by those Göttingen University collections in which donations of Baron von Asch are preserved. As it turned out (and we had expected), the data formats are as diverse as their providers and range from spreadsheets to proprietary database management systems. Furthermore, a variety of vocabularies, home-made thesauri as well as authority files, are used for the description of resources. Formatted for human consumption, a large part of the resource descriptions examined in the collections is not in an appropriate condition to be accessible and reusable with regard to a semantic interlinking on the Web. A transformation of the stored knowledge into sharable data available for a wider audience would be only feasible by structuring the data compliant to appropriate metadata standards. Therefore, the key elements of the model, its classes, properties and the relations between the entities, described allusively in the existing data, had to be identified.

3.2. Step Two: Formulating of Use Cases

The results of the workshop showed the heterogeneity of entities and relations needed for describing resources in the different disciplines, but also the similarities. It became apparent that each discipline needs its specific metadata scheme to describe their objects but that some components of the description should be reusable in other disciplines. Especially the reuse and interlinking of provenance information was seen as a step forward because it allows the contextualization of objects examined in different disciplines but once present at the same event (e.g. when a book and a tool were bought at the same time at the same place by the same person). Another aspect broadly discussed in the focus groups was how to verify the reliability of statements by metadata descriptions or interlinking with evidence.

The results were affirmed by the work and research experiences of the experts we interviewed. We clustered the reports into dimensions of topics, needs, entities, and relations and anonymized, generalized or specified the statements. Using this material we formulated case studies taking into account the recommendations of the cultural heritage aggregators DDB[9] and Europeana[10]. The short hypothetical stories of the case studies reflect research activities, describing the usage of

[9] https://pro.deutsche-digitale-bibliothek.de/teilnahmekriterien

[10] http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Mapping_Guidelines_v2.3_042016.pdf

metadata by users in general as well as in different contexts. These case studies were differentiated into scenarios, smaller units describing a specific usage of metadata by a user. We focused on those scenarios that many of the disciplines had in common, e.g. (1) "a user needs information about all events in the lifecycle of an item during a certain time span" (Scenario 22) or (2) "a user needs information about items that were present during an event" (Scenario 23). One or more scenarios describing individually the actions and aims of an actor were specified in each use case. Requirements were gathered from these scenarios describing the rules and constraints necessary for the realization of an action. The following use cases relevant to further work were figured out:

- **Information about resources,** described by scenarios related to the search and finding of resources on the Web in general;
- **Identification of resources,** described by scenarios related to the identification of resources;
- **Information about the history or lifecycle of resources,** described by scenarios related to information about events or activities the items of a collection were involved;
- **Information about change of use and reception of resources,** described by scenarios related to the change of use or reception an item underwent in its history;
- **Proof of information by evidence,** described by scenarios related to the description or search of evidence that proof the reliability of information;
- **Reliability of statements**, described by scenarios related to information about the description of statements by statements;
- **Access to resources**, described by scenarios concerning the usability of resources; and
- **Reuse of data,** described by scenarios related to the use of metadata descriptions by others (for more information see http://asch.wiki.gwdg.de/index.php/Use_Cases).

3.3. Step Three: Analysis of Requirements

The use cases, consisting of one or several scenarios, helped to achieve an abstract level by analyzing the possible interactions between an actor and a system. The fundamental structure of a scenario is composed of an identified actor and one or several identified goals this actor is pursuing. With the object to gain an identified goal, one or more requirements can be necessary or even be mandatory. E.g. the above-mentioned Scenario 22 is connected to two requirements: (1) Requirement 20 (Item descriptions must be interlinked with 1-n events in the lifecycle of the item) and (2) Requirement 27 (An event in the lifecycle of an item must be related to 0-n date information). In order to organize the gathered material, we subdivided the requirements according to three aspects:

- Requirements concerning the end-user: One of the determining factors for modeling a scheme is the context of usage it shall apply to. Therefore, it is indispensable to examine the field of use, the target group, and the language of data that would be appropriate.
- Requirements concerning the metadata: It has to be analyzed which properties of entities and what relationships between these entities must be taken into consideration.
- Requirements concerning the system: The functional settings of the system have to be examined because they are responsible for the accessibility to the data as well as for its representation and retrieval.

All in all, we figured out about eighty requirements.

As we defined the class Resource to be the superclass of all classes used in the ASCH model, the requirements concerning this class would be valid for all subclasses. More detailed and specified requirements were additionally assigned to the subclasses. Within the framework of this paper we can list some examples only (1) for superclass: e.g. resource descriptions must be machine readable, resource descriptions must be compliant to the one-to-one principle, resources must be interlinked with each other using unique, machine readable and persistent identifiers; (2) for the subclass Event, for example: e.g. an event in the lifecycle of an item must be related to 1-n items, an event in the lifecycle of an item must be related to 0-n places, an event in the lifecycle of an item must be related to 0-n date information. A detailed documentation is to be found in the ASCH wiki (see http://asch.wiki.gwdg.de/index.php/Use_Cases).

Concerning the provision of data, yet another aspect is significant – according to the Semantic Web and Linked Data, an application should not only be restricted to the concrete requirements of individual end-users, it also should keep a close eye on the possibilities of the networking opportunities given within the WWW. It is precisely for this reason that requirements, resulting from "metadata standards" used by the target communities, are taken into consideration when a certain profile shall be developed (see Zeng; Qin, 2008). Therefore, classes used in the ASCH model were aligned to classes from metadata schemes commonly used in the cultural heritage world.

3.4. Step Four: Identification of Classes and Relations between Classes

According to the requirements elicited from the use cases and scenarios, and considering the range of classes applied in common metadata schemes relevant for the description of collection items and their provenance, following classes (i.e. the superclass Resource and twelve subclasses) were identified to be used in the ASCH model:

- **Resource:** The superclass of all classes used in the model, all requirements valid for this class are also valid for all other classes of the model.
- **Metadata set:** The machine-readable description of a single resource represented by statements.
- **Item:** A real world thing in a collection.
- **Evidence:** A resource proving the reliability of a statement about a resource.
- **Event:** An activity in the lifecycle of a resource.
- **Time:** A time-span related to a resource via an activity or as a topic.
- **Agent:** A person, organization or group related to a resource via an activity or as a topic.
- **Place:** A geographic location related to a resource via an activity or as a topic.
- **Digital representation:** A digital resource depicting an item.
- **Collection:** An aggregation of items.
- **Statement:** A predication about an item.
- **Holding:** The place an item is located.
- **Concept:** A term from an authority used as a value in a resource description.

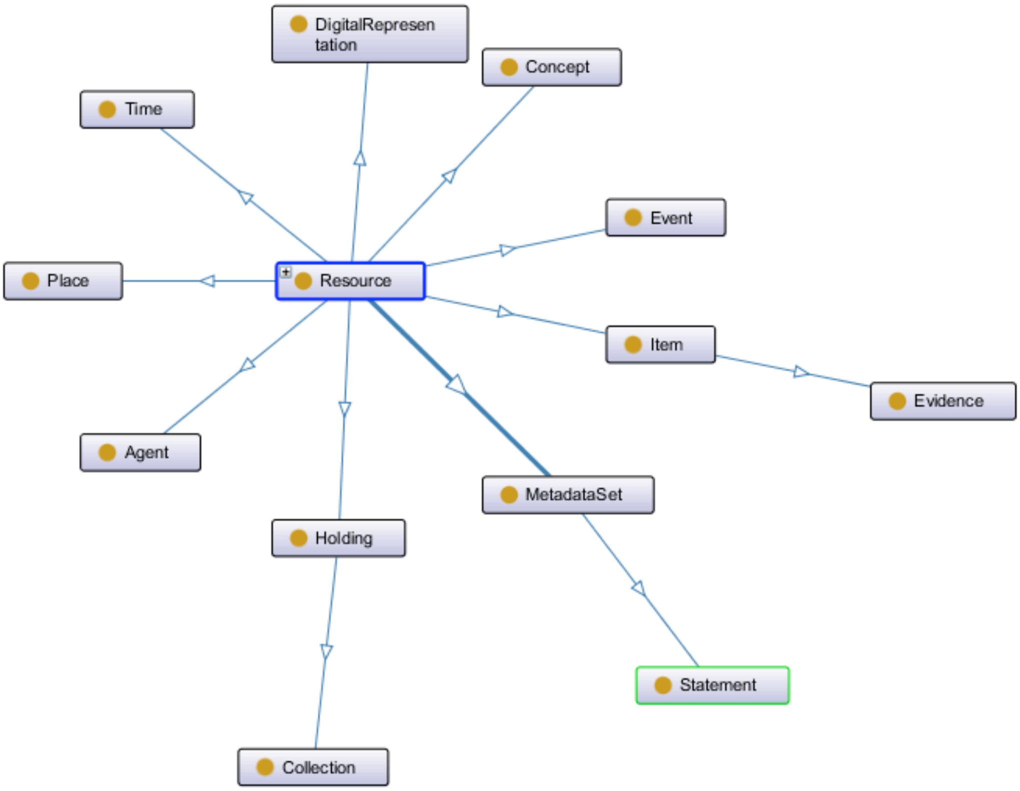

FIG. 1. The ASCH model classes and relations.

As the project focuses on the standardized description of the semantic contextualization of objects and especially provenance information about these objects, we considered only those standards relevant that are appropriate for a semantic contextualization. At present, these are Dublin Core (with DC Metadata Element Set and DCMI Metadata Terms), EDM, PROV-DM, CIDOC CRM, and Darwin Core (DwC), an extension of DC for biodiversity information. Scope of the selection of these standards is interoperability and the cross-domain use of metadata. In the DCMI Glossary[11] the term 'interoperability' is defined as "the ability of different types of computers, networks, operating systems, and applications to work together effectively, without prior communication, in order to exchange information in a useful and meaningful manner." Developed for different scientific and institutional fields, these standards are focusing on diverging requirements. DCMI e.g. provides standards especially for a generic description of resources on the Web, but is also a Linked Data compliant standard. PROV is developed as an RDF standard for the description of provenance information of web resources, leaving a further description of the resource to other standards. CIDOC CRM, a semantic model that forms a base for other metadata standards (e.g. LIDO or EDM), is concentrating on the events in the lifecycle of items and DwC allows the detailed taxonomical assignment of items. RDF as one requirement to make data linkable will be used with terms of the chosen standards and evidence shall be interlinked with object descriptions because one of the requirements relevant for the research community is the interlinking of metadata descriptions of objects with parts of text in evidence

[11] http://www.dublincore.org/documents/usageguide/glossary.shtml

encoded in TEI[12] and referencing to these objects. In this context it will be possible to describe who made which statement when and where, and how reliable a statement is.

Figure 2 illustrates the provenance component of the ASCH model. All classes in the ASCH model will be identified as subclasses of the above listed RDF compliant ontologies. So an item of a zoological collection may be described using DwC and thereby be compliant with other data from the same discipline. Then the provenance description via the Event class is using DwC properties and classes parallel to PROV properties, and classes where the PROV properties and classes are a hub that allows the contextualization of this data with data from other disciplines describing the same event.

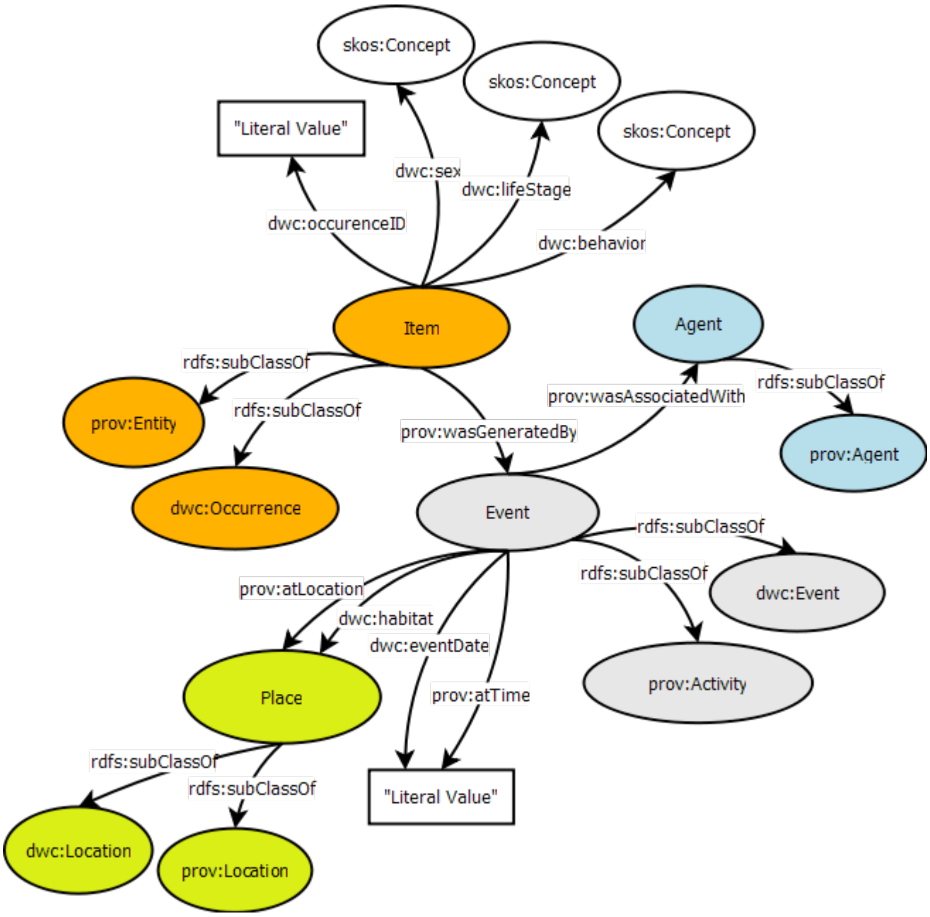

FIG 2: Description of zoological items using DwC and PROV Ontology

4. Conclusions and Future Work

To help making the huge amount of digitized cultural heritage objects accessible, it is indispensable to align metadata about these objects using a hub for those components of the description that are relevant across disciplines. This would clear the way to the interoperability of data across the borders of the various disciplines because it allows the use of domain specific metadata terms where necessary and common used terms where possible. To find out which components are usable for such a hub, we discussed the differences and similarities with experts working in natural science and humanities and in different cultural heritage institutions. The

[12] http://www.tei-c.org/

result was an abundance of information allowing us to identify those scenarios and requirements relevant for all experts independent from their background and research environment. Based on these results we started to identify the relevant classes and aligned them to RDF compliant metadata schemes used in the different domains. With the definition and alignment of these classes we finished the first four steps of our project.

Step five of the project's working plan we have already started to work on is the "identification of properties". The procedure was similar to that used for the identification of classes. According to our scenarios and requirements we initially defined the needed properties in a form independent from a common metadata standard. Then we started to align these properties to properties from the RDF compliant schemas listed in chapter 3.4. Properties and classes will then be used to develop domain specific application profiles and profiles for the hubs in step six. The abstract representation of interlinked entities and the characterization of the various relationships in the model will turn into substantiality by testing the model's functionality with concrete data describing objects known to have the provenance Baron von Asch. The tests will be carried out in various annotation systems and are defined as the last step on our way to develop the ASCH model.

References

Baca, Murtha (ed.; 2008): Introduction to Metadata. Los Angeles: The Getty Research Institute.

Berners-Lee, Tim (27 July 2006): Linked Data Design Issues. 27 July 2006. W3C-Internal Document. http://www.w3.org/DesignIssues/LinkedData.html.

DiNucci, Darcy (1999): "Fragmented Future". In: Print 53,4; pp. 32, 221.

Gilliland, Anne J. (2008): "Setting the Stage". In: Baca, Murtha (ed.): Introduction to Metadata. Los Angeles: The Getty Research Institute; pp. 1-19.

Hauser-Schäublin, Brigitta; Gundolf Krüger (eds.; 2007): Siberia and Russian America: Culture and Art from the 1700s. The Asch Collection Göttingen. München, Berlin, London, New York: Prestel.

O'Reilly, Tim (09/30/2005): What is Web 2.0. O'Reilly Media. http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. Retrieved 2016-08-19.

W3C (2015): Semantic Web. https://www.w3.org/standards/semanticweb/. Retrieved 2016-08-19.

Zeng, Marcia Lei; Jian Qin (2008): Metadata. London: Facet Publishing.

Zeng, Marcia Lei; Jian Qin (2008): "Schemas – Structure and Semantics". In: Metadata. London: Facet Publishing; pp. 87-130.

Wood, David; Marsha Zaidman, Luke Ruth with Michael Hausenblas (2014): Linked Data. Structured Data on the Web. Shelter Island, NY: Manning.