Date submitted: 22/06/2010

Linked Data for Libraries

Jan Hannemann
E-mail: j.hannemann@d-nb.de

and

Jürgen Kett
E-mail: j.kett@d-nb.de

German National Library,
Frankfurt, Germany

Meeting: **149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

The Semantic Web in general and the Linking Open Data initiative in particular encourage institutions to publish, share and interlink their data. This has considerable potential for libraries, which can complement their data by linking it to other, external data sources.

This paper details the first linked open data service of the German National Library. The focus is on the challenges met during the inception of this service. Extrapolating from our experiences, the paper further discusses the German National Library's perspective on the future of library data exchange and the potential for the creation of globally interlinked library data. We outline how this process can be facilitated and how new services can be offered based on these growing metadata collections.

1. Introduction

Libraries are currently largely isolated in terms of data exchange, since data is primarily collected by libraries for libraries. The process of exchanging and jointly utilizing data with non-library institutions is still in its infancy. Collaborations exist primarily between libraries only and library data is not yet an integral part of the web. This is mostly due to the poor level of linking between library datasets and data from other domains, but also due to the current data collection processes and data formats, which – naturally – focus on classical usage scenarios for libraries.

The Semantic Web and especially the Linking Open Data initiative encourage institutions to publish, share and cross-link their data using the web. Data visibility can vastly improve through interlinking with other information sources. This is relevant both for non-profit and commercial institutions. Becoming part of the linked data web, or "semantic cloud", also means that libraries can better meet user expectations, such as continuous availability of information in a format that is understandable by non-library-experts. Working within the growing knowledge base of the cloud can also help with many complex tasks that libraries are currently confronted with when they maintain and optimize their own local datasets. Important examples for those tasks are duplication detection, disambiguation, individualization, data quality management and enrichment. It further paves the way for new services that utilize more than a single institution's data. The linked data community, in turn, also benefits from such efforts by libraries and other cultural heritage institutions. Library data tends to be of very high quality, being collected, revised and maintained by trained professionals. As such, it has the potential to become a much-needed backbone of trust for the growing semantic web.

Libraries have recognized this potential, and several institutions are planning to publish their data as Linked Data. In practice, however, this is a challenging process. In addition to the organizational hurdles, the technical side of publishing data for and using data from the semantic web can pose a considerable problem for cultural heritage institutions like libraries, especially for those with a more limited IT budget.

The purpose of this paper is to discuss linked data from the perspective of libraries and other cultural heritage institutions, as well as to provide a concrete report on the German National Library's experience with establishing such a service.

2. Vision: The global cultural graph

The main problem for the linked data web is dealing with reliability: Is the data correct and do processes exist that guarantee a high data quality? Who is responsible for it? Of the same importance is reliability in time: Is a resource stable enough to be citable, or will it be gone at some point? These questions are of special importance in the context of research, where citability is essential, and for higher-level services that are based on this kind of data.

While it is not necessary for every dataset to provide a maximum degree of reliability in order to be useful, we believe that in the heart of the linked data web a stable core is needed, a backbone of trust, and that cultural heritage institutions are in a unique position to provide parts of this core: connecting the local knowledge bases of all cultural heritage institutions could lead to a huge *global cultural graph* of information that is both reliable and persistent.

To account for different degrees of data quality and reliability, the data of the global cultural graph should be organized in a shell model (see Figure 1). Each layer would be associated with a policy that is getting stricter towards the reliable core of the knowledge base. Entities located in the core are persistent and therefore reliably citable. The descriptions of these entities should be versioned so that changes in the statements and their provenance are persistently documented. Transparent policies will be used for each of the layers that are based on well-documented, accepted standards and cataloging rules. In order to guarantee data quality and persistence, data in the core must be backed by one or more trusted public institution.

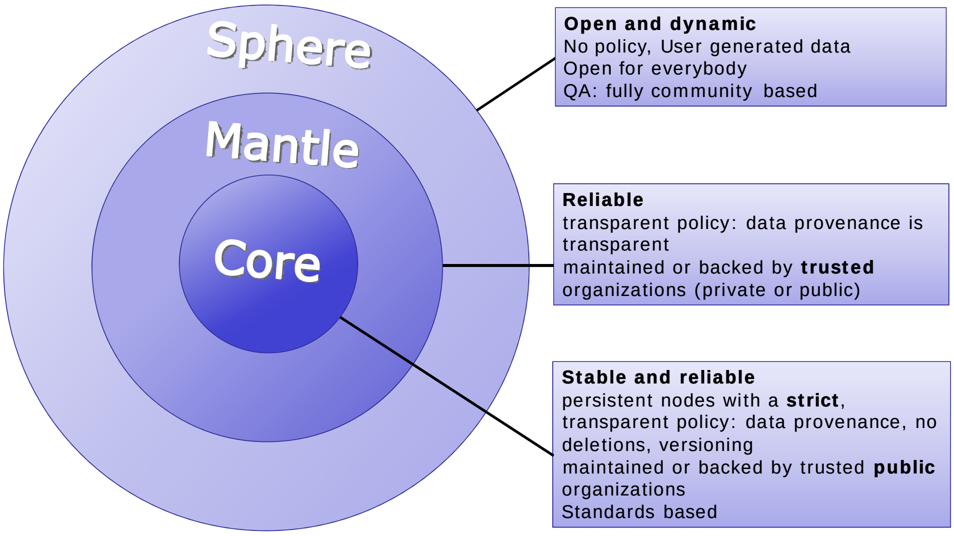
Sphere
Mantle
Core
Open and dynamic
No policy, User generated data
Open for everybody
QA: fully community based
Reliable
transparent policy: data provenance is transparent
maintained or backed by **trusted** organizations (private or public)
Stable and reliable
persistent nodes with a **strict**, transparent policy: data provenance, no deletions, versioning
maintained or backed by trusted **public** organizations
Standards based

Figure 1: Shell model of reliability and persistency

This model may include even automatically generated metadata, as long as the provenance of the data is documented. Essential is the possibility for information to move towards the core if it meets the required rules. That way, the core of trusted information, as well as the value of the entire data aggregation, can grow over time.

The conditions for realizing this vision are promising: cultural heritage institutions already use collaboratively developed and well-documented standards, such as MARC21 or RAK-WB, and rules to produce and maintain their data, even if they would need to be adjusted to be suitable for general data exchange on the web. In the German library community the practice of data exchange and cooperative maintenance of central databases has led to workflows that form an excellent basis for providing stable datasets for the web. A good example for that are the cooperatively maintained German authority files for persons, corporate bodies and subject headings (PND, GKD, and SWD, respectively). Well-defined processes for deletions, updates and the merging of duplicates, together with an already established identifier scheme guarantee a high degree of stability. Therefore, we chose these data sets for establishing our own first linked data service.

Publishing our local knowledge bases as linked data is an essential step towards the vision described above. In the following, we highlight the challenges involved in this step and the experiences we made.

3. Challenges in setting up a linked data service

Despite common claims to the contrary, setting up a linked data service is not trivial. Cultural heritage institutions interested in becoming part of the cloud have to deal with a variety of challenges, which can be grouped into the following categories.

3.1. Technical

In order to set up a linked data service, a certain amount of infrastructure is required. Generally, this comprises a means of data storage (typically a triplestore or a database), a web server, and a resolver that interprets incoming web requests, translates them into queries for the data storage, and returns the results.

Given the relative novelty of the linked data movement, the technological options are still largely in development – and often poorly documented. In particular, it is often unclear to institutions new to linked data which of the many options is best suited for their purposes.

3.2. Conceptual

Another essential issue is that of data modeling. There are dozens of more or less widely established ontologies to choose from, each one with a set of advantages and drawbacks. An important aspect to look out for in this context is the definition of individual properties, which may or may not fit the data to be modeled. If no ontology can be found that is ideally suited, it might be necessary to mix ontologies and/or to extend them with custom properties.

Certain information is particularly difficult to model, such as statements about statements. Examples for that are specifying the provenance of a certain statement or about the processes and rules that where used to create the data. The latter point is of special importance for data that has been produced by automated algorithms, but is also necessary to document the manual cataloguing rules and standards being used. Various approaches exist to address this problem, such as N-ary relations[1], OWL 2 axiom annotations[2], reification[3], quads and custom ontology extensions. Each approach comes with its own set of advantages and drawbacks. There are no established best practices in this area and no community consensus on which approach is best suited for libraries. Essentially, experiences with modeling library data using these approaches is still missing.

Another issue is the specification of URIs. Typically, (non-library) organizations that publish their data own an isolated dataset without any public identifier and existing data exchange workflows. Introducing new URIs for entities and their descriptions is therefore not an issue. Libraries, on the other hand, already use lots of public identifiers for their data and the entities they describe, and libraries are massively exchanging data. We believe it is better not to separate the world of linked data and the traditional data exchange procedures, but rather to join them. A suitable identifier scheme should work for all workflows - identifiers should not vary depending on the data exchange protocol being used or the data representation (e.g. RDF versus MARC21).

[1] www.w3.org/TR/swbp-n-aryRelations

[2] www.w3.org/TR/owl2-syntax/#Axioms

[3] en.wikipedia.org/wiki/Reification_(computer_science)

3.3. Legal

In terms of legal matters, two issues are important: publication rights and licensing of linked data. Cultural heritage institutions like libraries tend to collect data, oftentimes in cooperation with other institutes. In such cases, it has to be clearly established what data can be published as linked data, i.e., made publicly available.

For both bibliographic and authority file data, privacy issues may be a concern. For example, if a library collects information about authors, it is possible that individual authors do not want their personal information (birthdate, places of birth or residence, association with corporate bodies, etc.) published.

Licensing issues can be a problem as well. Terms of use for linked data should be decided upon early, as legal checks may take time. It can be helpful to distinguish between commercial and non-commercial use, with the latter one being often offered for free.

3.4. Overall

A common issue that touches all three categories above is the general lack of experience reports for establishing linked services (something this paper intends to address). Step-by-step instructions are very rare and generally leave many questions unanswered. A variety of best practices exist, but while offering the much-needed flexibility for institutions intending to become part of the linked data web, they also provide less of a guideline than strict standards could.

4. DNB Linked Data Service

This section describes our experiences with establishing a first linked data service. The project was kept intentionally small to avoid over-planning and the setting of unrealistic goals. Similarly, we decided early on that we would focus on selected parts of our data collections instead of providing a comprehensive solution for all our data.

For the technical and conceptual development necessary, we chose an iterative approach with each cycle lasting 1-2 months and culminating in a successively more elaborate working prototype. This allowed us to get early feedback from the community and potential users and incorporate it into the design of the final service.

4.1. Technical realization

The system architecture is shown in Figure 2. All our library data is stored in a central database, where it is constantly updated and added to. To make parts of this data available as linked data, a component called RdfExporter extracts the metadata from our central cataloging system[4], converts it to RDF and stores it in a Jena TDB[5] RDF Store. In a second step this data is enriched with references to external sources. The required data transformation is performed by a set of conversion modules (one for each kind of data, i.e., separate modules for persons, corporate bodies and subject headings) that are provided by our central Conversion Service. A Joseki Server allows access to the data.

[4] OCLC CBS: oclc.org/cbs/default.htm

[5] openjena.org/TDB/

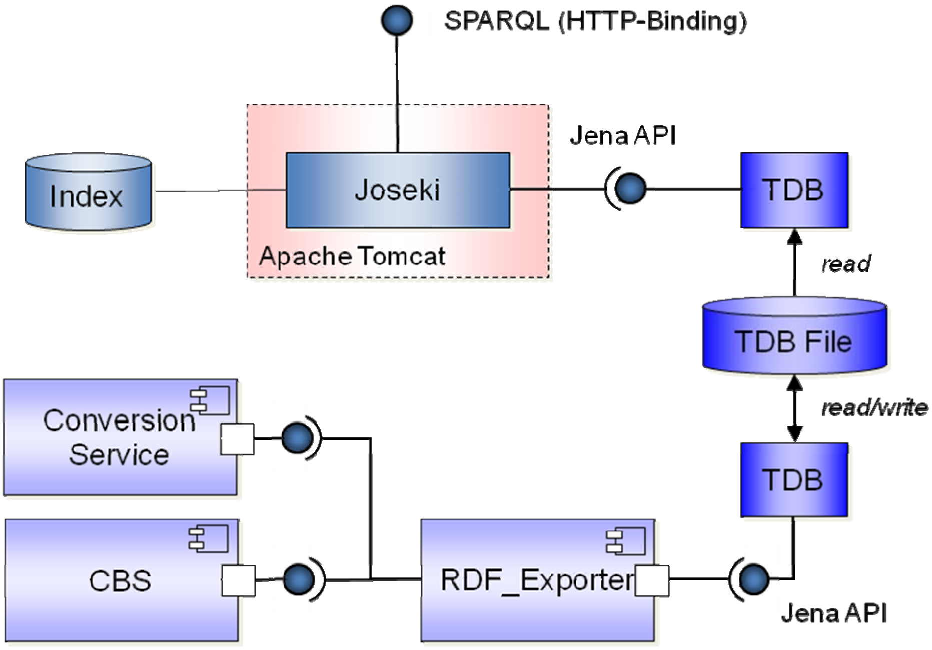

Figure 2: Current system architecture

We have chosen these particular technologies for data storage and access since we have used them in another project before. However, it has to be pointed out that they are still in active development and the documentation is consequently lacking depth.

The current solution works, but will likely be replaced in the future. Scalability, in particular, is an issue we need to address. Already the large amount of data involved poses a challenge: a complete conversion of our three data sets takes about two days on a modern computer, and that neither includes adding external links nor joining the enriched individual datasets together. The Joseki server also has a very large memory footprint (several gigabytes, growing with the size of the database) and we have not found a way to reduce it yet.

Another issue is that the current solution requires the data transformation to be initiated manually, which is one of the things we will address in the future. Ideally, the data conversion and enrichment with external links would happen on demand (i.e., when a request is received) or associated with an automatic update mechanism, which is triggered whenever the CBS database changes in a meaningful way.

4.2. Data selection

An important question to ask when setting up a linked data service is what data is going to be published and which external data sources are going to be linked.

We decided to focus on authority file data, in particular information about 1.8 million persons (from the Name Authority File PND), 160,000 subject headings (from the Subject Headings Authority File SWD), and 1.3 million corporate bodies (from the Corporate Body Authority File GKD). This choice is mostly due to external requests for said data, the manageable scope and the data’s properties (see also section 2):

- The data is already used by many organizations.

- The maintenance workflows for this data are suitable for publishing and guarantee a high degree of persistence.
- An established and accepted identifier scheme for these datasets already exists.
- In terms of links to external data sources, we were in the lucky position to be able to draw on the results of several projects and cooperations between libraries that align data sets. We can offer links to the German Wikipedia[6] and DBpedia[7], to VIAF[8], LCSH[9] and RAMEAU[10].

Our bibliographic data is much more extensive, which is why we will address it in a later project.

4.3. Ontology Selection

Selecting established ontologies as the basis for data modeling is strongly suggested in the semantic web community, since it makes the published data easier to share and exchange. Consequently, we aimed to do that as well.

In practice however, we had to realize that existing ontologies are only partially suitable to model our data. Individual properties had definitions that did not match our data sets, so that no single ontology was found acceptable. Instead, we had to meticulously determine a set of ontologies whose parts would together cover most of our data. For the remaining portions we defined our own properties, with the intent to register the resulting ontology in the future.

The data modeling for the representation of persons and corporate bodies utilizes several existing ontologies like the RDA element sets, the FOAF vocabulary and the Relationship Vocabulary, with RDA being the basis for our data modeling due to the fact that it covers the relevant Functional Requirements for Bibliographic Records (FRBR) entities (i.e., persons and corporate bodies) very well. We further used properties created by the German National Library (Gemeinsame Normdatei (GND) vocabulary) to complement these ontologies. For subject headings the data modeling is based on the use of the Simple Knowledge Organization System (SKOS) and Dublin Core elements, which are also complemented by special GND-properties.

A detailed discussion of the various ontologies we considered and the reasoning for our final selection of ontology elements is part of the documentation of our linked data service[11].

[6] de.wikipedia.org

[7] wiki.dbpedia.org

[8] viaf.org

[9] authorities.loc.gov

[10] rameau.bnf.fr

[11] wiki.d-nb.de/display/LDS

4.4. Examples

The following examples illustrate the work completed in the project:

- The German author *Bertolt Brecht's* (http://d-nb.info/gnd/118514768) XML/RDF representation is found here: http://d-nb.info/gnd/118514768/about
- The XML/RDF representation of the corporate body "*IFLA / Section of Public Libraries <The Hague>*" (http://d-nb.info/gnd/10352988-3) is located here: http://d-nb.info/gnd/10352988-3/about
- The subject heading for "*Führungskraft*" (English: "*Executive*") is found here: http://d-nb.info/gnd/4071497-4, and the associated XML/RDF representation is located here: http://d-nb.info/gnd/4071497-4/about

5. Experiences

We have been involved with semantic web activities for a while before we decided to establish our own linked data service. Therefore, we thought us well-prepared for the challenges involved. Our practical experience, however positive, also showed several pitfalls and somewhat misguided expectations. In particular, the popular claim of the semantic web community that setting up such a service is easy has to be taken with a grain of salt. Our findings were the following:

- *Setting up a service is not trivial.* Linked data initiatives are a rather recent development. As a consequence, the essential software solutions (tools) involved have not reached full maturity yet. That means, among other things, that documentation may be lacking the required depth. For a working service, multiple software components need to be setup to work together (e.g., see Figure 2), which requires appropriate expertise. The data will likely have to be transformed into a suitable format (RDF). This part not only requires a proper modeling of the data/transformation (a potentially considerable effort if the data modeling is to match the original data closely), but also the creation of conversion programs or export filters. For the final data representation UTF-8 encoding should be used, even if libraries use a different encoding internally.
- *Data modeling can be complex.* When publishing data on the web, it is advantageous to use existing, registered ontologies. Unfortunately, these ontologies do not always match the data representation of each individual library (see section 3). In particular, the definitions of individual properties can vary considerably. There are two general ways to deal with this issue: either the published ontologies are simply taken as-is, or new properties must be defined to match the data. The former approach is easier, but may distort the data, while the latter is much more complex, but represents the data properly. There is no simple answer to the question which is the right thing to do. We chose the model to match our data as closely as possible, since we believe that it would compromise our data quality to do otherwise.
- *Open data exchange mentality does not exist everywhere.* Even before linked data, libraries have exchanged and aligned their data sets. The results of such projects could be prime information sources for connecting linked data sets. Sadly, not all institutions involved share the open exchange mentality, and shared

ownership may make it difficult to publish these results. We've made both positive and negative experiences in this regard, and recommend that libraries in a similar position discuss the matter with all involved parties well before considering using such collaboration results.

- *Best practices are seen as rules.* Linked open data is based largely on best practices rather than rules. However, this pragmatic aspect is not seen as essential in all areas of the linked data community. Deviations from perceived standards tend to be criticized, which can cause institutions new to the semantic web to doubt their decisions – even if they make sense for the organization in question. Libraries should not be deterred by such feedback and rather see this as a motivation to contribute their own experiences and knowledge to the community. Guidelines and best practices should be carefully considered in the context of each institution's needs, especially in this early forming phase of the semantic cloud. For example, we have been criticized for not offering a SPARQL-endpoint. While it certainly is a useful addition, it is not entirely clear why this tends to be seen as a must-have instead – especially in the light of existing and established alternatives for the searching (e.g., Search and Retrieval via URI[12] (SRU) and OpenSearch[13]) and synchronizing (e.g., ORI[14]) of library data.
- *Users remain largely anonymous.* To improve our service, we asked ourselves two fundamental questions when we started this project: who is using our data? And: what is the data used for? Although we do invite users to comment on our service and to tell us of their expectations and experiences, most users choose not to. It is a consequence of the linked data concept of anonymous access that we are only in touch with those users who do contact us. This also means that we cannot provide specific help to other users: an odd development we observed was that someone apparently wrote a program to crawl our linked data using a (very) naïve approach (trying all possible numeric combinations that might constitute an IDN to generate URIs); this program will run for many months before completing. It would be much easier to direct that user to the downloadable version of our data that we provide, but since the user is using a dialup connection, we cannot identify and contact him/her.
- *Properly modeled data is very useful.* Once the data modeling is completed and the data made available, it can be used by others. A colleague at the Technical University of Braunschweig has shown that with properly modeled data, this can result in very useful applications: within a day, he imported our data into a database, added a web interface and had thus created a searchable access to our data.

Overall, we can draw a positive summary despite the challenges outlined above. Once the hurdles are taken, it is relatively easy to utilize the data and to expand the service.

[12] www.loc.gov/standards/sru/

[13] www.opensearch.org

[14] www.openarchives.org/OAI/openarchivesprotocol.html

6. Future Work

In establishing a linked data service, we have taken one step towards the vision of a global cultural graph – but there are still many steps to take, and the long term goals are only achievable when other institutions join in. Cultural heritage organizations as a whole have to adapt their technical infrastructure, business processes, rules, policies and licenses for their metadata to meet the requirements of the web. This fundamental change requires a stepwise and manageable approach. First and foremost, the challenges for cultural heritage institutions have to be addressed. Once the cloud is growing, global strategies can be developed.

In this section, we show our own plans for the future, as well as higher goals that we can only achieve in cooperation with other institutions.

6.1. Short-term goals

- *Revised Infrastructure:* our immediate goal for the future is to improve our current service both in terms of its infrastructure and data representation and to add new datasets. For the next version of the service, we will provide automated update mechanisms (see above) for our data and potentially for links to other data sources as well. A more scalable architecture will have to be found, too. To provide a broader access to our data, we will provide an SRU interface.
- *New datasets:* we will add classifications and improve the modeling of the data we already provide. The new datasets will contain, among others things, a selected subset of the German national bibliography and the German translation of the Dewey Decimal Classification. For the title data a decision has to be made regarding an URI-Scheme and a registration workflow is needed that is also compatible with the current plans of optimizing exchange and reuse of bibliographic data in Germany. We are currently discussing this issue jointly with German library centers and hope to reach an agreement soon. Another big issue will be the ontology modeling for title data. Again we have to find a compromise between re-using existing vocabularies like Dublin Core[15] and the bibliographic ontology[16], considering evolving standards like RDA and our existing data structures.
- *End-user services:* one reason for providing our data on the web is to make it possible for us and others to build richer services utilizing this data. Cultural heritage institutions should not be the only ones offering services for their data; the creative potential of the web community can accomplish more than individual institutions alone. Our goal is to provide at least one reference implementation of a service that illustrates the full potential of interlinked library datasets. The service addresses two target groups: it shall motivate other cultural heritage institutions (especially in Germany) to contribute to the global cultural graph and it is intended to attract the attention of the web community to these valuable datasets. The service will resolve a URI that identifies a certain entity (e.g. book, work, person, corporate body, topic, etc.) and return an information site with links

[15] http://dublincore.org/documents/dc-rdf/

[16] http://bibliontology.com/

to all registered datasets that provide (directly) related resources. This service will follow the linked data conventions and offer an RDF representation of the entities.

6.2. Long-term goals

It is obvious that the long-term goals are only achievable when cultural heritage organizations, software vendors, research institutions and public authorities are acting in concert. The German National Library will use its influence and technical competence to drive this development forward together with her partners. The following issues are those we identified to be most urgent:

- *Shared Licensing Model:* libraries are indexing more and more in cooperation. The creation of the Deutsche Digitale Bibliothek[17] will have German libraries, museums and archives grow even closer together. In a world of jointly generated metadata, appropriate licensing models can only be cooperatively developed and established. If that is not possible, reuse of the data in the cloud will be severely limited, to the point that we would need to handle data licenses not on the record level but on the level of isolated metadata fields (e.g., if one institution enriches another institution's data). In practice this would be far too complicated and not cost effective.
- *Adopting workflows and policies:* cultural heritage organizations should adopt their workflows to make their datasets less redundant and more persistent (especially title data on the Work and Manifestation level). There is a need for citable public identifiers that we all agree on and we all will use. The German *Gemeinsame Normdatei* (GND) could serve as a model for such initiatives.
- *More proved design patterns for ontology modeling:* cultural heritage institutions can not solve the fundamental challenges tied to data modeling (e.g., statements about statements) and technical realization themselves, but they can provide practical use cases. Software vendors and research institutions are needed to develop suitable solutions for these problems.
- *Advanced technical solutions:* the current state of the art regarding technical solutions is not suitable for the productive environments of cultural heritage organizations. Tools and frameworks that facilitate releasing data into the cloud and utilizing the data sources already available are needed. Linked data technology must seamlessly integrate in the existing environments and workflows. Therefore, new developments should be platform-independent, follow open standards, and should use a highly modularized layered architecture with open APIs. At the same time, existing library systems have to evolve in parallel: in the future they must be able to handle granular provenance and versioning information.

[17] www.deutsche-digitale-bibliothek.de (in German)

7. Summary

Linking open data offers a great potential for cultural heritage institutions like libraries. Not only can data see much more widespread use through it, but by cross-linking sets of data together, the value of the resulting data cloud grows. Based on it, libraries and other organizations can create novel and more comprehensive services. In establishing a linked data service, we have taken one step forward towards this vision, but the long term goal of a *global cultural graph* is only achievable when this approach for data publication and exchange receives widespread support.