http://conference.ifla.org/ifla78 Date submitted: 7 June 2012

# The Deutsche Nationalbibliografie as linked open data: Applications and opportunities

**Jürgen Kett**

**Sarah Beyer**

**Mathias Manecke**

**Yvonne Jahns** and

**Lars G. Svensson**
Deutsche Nationalbibliothek
Frankfurt am Main, Germany

**Session:** 215 — *What is a national bibliography today and what are its potential uses?* — Bibliography

**Abstract:**

*This paper discusses the requirements on a national bibliography in the 21st century. Building on the traditional criteria data completeness and reliability, data currentness, referentiability, and persistence, the authors add the criterion that a national bibliography needs to integrate itself into the World Wide Web, since this is where information exchange takes place today. This should be made using linked data technologies and the data should be published under an open license. As a case study, we present the work done at the Deutsche Nationalbibliothek, where circa 70% of the database is already published as linked open data.*

## Introduction

A national bibliography can be defined as the complete list of publications from a geographically restricted area.[1] In the age of the internet, however, it is not trivial to carry this definition forward, since the internet changed both the meaning of "publication" and of the way we publish.

Where a traditional printed book is self-contained and thus static, there are internet publications which have no well-defined borders and that can be dynamic and interactive, containing ever-changing content. Further, the technical means to record the textual output have changed: with the possibilities of full-

[1] Cf. Anderson (1974) p. 12

text search, the use of metadata-based catalogues, thesauri, and classification systems have been marginalised. Not necessarily because full-text search is superior to metadata-based search, but because it is easier to automate: full text resources are readily available on the web, whereas high-quality metadata needs to be generated. Metadata creation – or cataloguing – requires either complex automated processes or human intervention.

But now much of this human-driven process is done outside of libraries: publishers use their own metadata to increase their visibility, and online platforms such as Wikipedia or the OpenLibrary invite the general public to publish articles and metadata describing publications, thus creating an ideal complement to the full-text search. The recent rise of mass digitisation and OCR changed the scene even more by opening the possibility to perform full-text search in (previously) non-digital material. For a search engine company like Google this obviously seemed more effective than to use traditional cataloguing methods – in their search engine, metadata-based search is merely seen as a complement.[2]

Against this background, libraries need to ask themselves which added value a national bibliography built on traditional cataloguing can deliver, now and in the future.

## Requirements for a national bibliography

Traditionally, the national bibliography targeted three different groups: publishers and the bookselling trade, libraries, and end users (particularly researchers and literary scholars). Those three groups had in common that they built on the following four basic properties of a national bibliography:

1) Data completeness and reliability
   For booksellers, publishers, and researchers it was and still is imperative that the national bibliography documented the complete (professional) output, without any political or contentual bias. Further, the adherence to cataloguing rules played a prominent role, particularly when constructing cumulations.
2) Data currentness
   Especially for the bookselling trade and libraries the currentness of the data in the national bibliography was highly important. In the 2nd part of the 20th century much effort was put into the data processing in order to ensure short publication cycles in spite of an increasing number of publications.
3) Referentiability
   Given completeness, reliability, and currentness, a national bibliography could serve as a reference point for scientific purposes: if a book was listed in the national bibliography, it was certain that it existed, and a book not listed there most likely had never been published.

[2] Cf. http://books.google.com/intl/en/googlebooks/about.html

4) Persistence
   In order to serve as a reference point for citations, it was not enough that the entries conformed to the above quality criteria, but it needed certain persistence, too. Up to the beginning of the 21st century this was not a problem, since the bibliography was a printed publication in itself. It was only possible to correct an improper bibliographic entry in a cumulation, but apart from that there was no way to delete the evidence that a book had been published.

The requirements for a future national bibliography need to go further and to measure itself on data use and re-use in the World Wide Web.

More and more the WWW mutates to an open space for data exchange: the so-called Linked Data Cloud.[3] Since 2008, this collection of interlinked datasets has increased enormously in size, but little is known about data quality and data persistence. In order for a semantic network to function, however, we need a certain level of reliability, both regarding information quality but also and particularly regarding information persistence. This network can only grow sustainably if we can know that the information we annotate today is available next week: it has to be citeable.[4] This problem applies in particular to online publications and their metadata. In order to ensure long-time reliability it is necessary to stabilise this part of the internet. This could be a future task for libraries and other cultural heritage organisations.

From this we can derive the following:

1) Data completeness and reliability
   The rise of online publishing makes it virtually impossible to collect alls creative output that is relevant for a national bibliography. Most likely, a national bibliographic agency will not even be able to reliably predict how incomplete their list of publications is. Instead, they will have to define for what kinds of publications want to achieve completeness (e. g. print media, high-quality blogs, and certain types of scholarly, self-contained publications) and where they are satisfied with periodic snapshots of a set of websites.
   Some metadata will be created by automated processes and will be less than perfect. This does not render the metadata useless, but the processes need to be well documented and it must be transparent to the data consumer that the information was machine-generated, and which are the suitable use-cases for that kind of data.
   Conformance to specific set of rules will need to focus less on cataloguing rules à la AACR[5] or RAK[6] and more on technical and semantic data interoperability for the integration in external services.

[3] For an introduction to Linked Data and the Linked Data Cloud cf. Heath and Bizer (2011)

[4] Cf. Schuster and Rappold (2006)

[5] Anglo-American Cataloguing Rules, cf. http://www.aacr2.org/

2) Data currentness
   Many data consumers expect that online publications will be listed in the national bibliography the moment it is published. This acts contrary to the requirements for referentiability and persistence: if the national bibliography is built on metadata describing early versions of a publication the risk that there will be changes in the bibliographic description is relatively high. This will require that library information systems are capable of archiving different versions of the metadata preserving different states.
3) Referentiability
   The national bibliography will in future keep its role as a central reference point for publications. Particularly for electronic publications, it can be expected that this role will be increasingly important. In order to provide this service, it is necessary to provide unique identifiers for bibliographic data that work in an electronic environment.
4) Persistence
   Long-term availability and ensuring the integrity of bibliographic descriptions are prerequisites for long-term referentiability. It is not only a matter of persistent identification of bibliographic and authority data, but also one of version control, handling of corrections and deletions, and identification of specific states, e. g. through the use of timestamps.

## The topology of the German National Bibliography in the digital age

Considering those requirements, it is obvious that a national bibliography must not separate itself from the World Wide Web. In Germany the national bibliography has transformed itself from a set of catalogue cards to an (electronic) database using more and more complex data structures, where the bibliographic descriptions link to authority data and often to resources outside the system. The next step is to let the national bibliography merge into the WWW, so that it in every aspect will be an integral part of the web: topologically, functionally, technically, and organizationally.

Looking at its characteristics, we could understand our national bibliography as a graph in the World Wide Web. This graph is interlinked with itself but also with other parts of the WWW. This view of library data as a graph in the WWW applies not only to the German national bibliography: other directories of cultural heritage objects, authority files, thesauri and classifications need to be part of the web, too.

[6] Regeln für die alphabetische Katalogisierung, the cataloguing code used in Germany and Austria, cf. http://www.dnb.de/EN/Standardisierung/Regelwerke/regelwerke_node.html#doc3132bodyText1

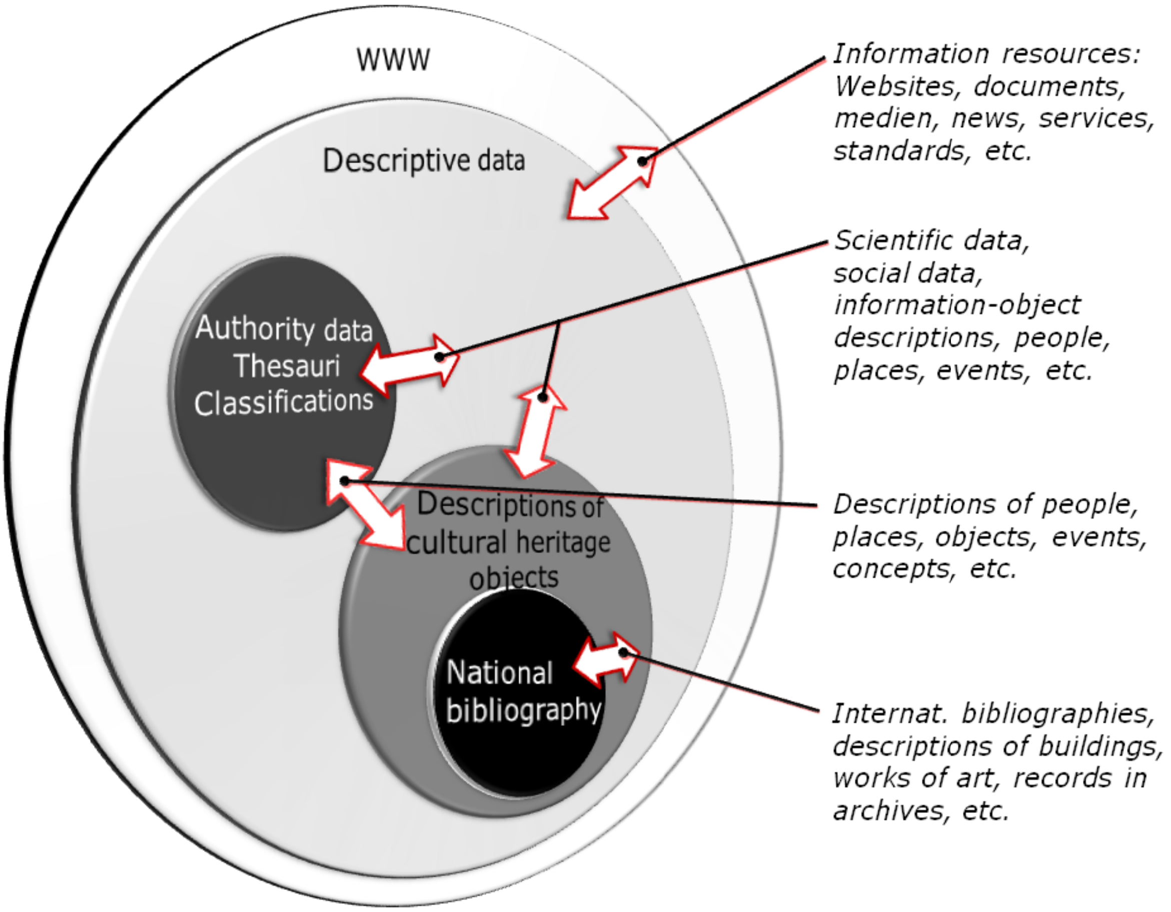
WWW
Descriptive data
Authority data
Thesauri
Classifications
Descriptions of cultural heritage objects
National bibliography
Information resources: Websites, documents, medien, news, services, standards, etc.
Scientific data, social data, information-object descriptions, people, places, events, etc.
Descriptions of people, places, objects, events, concepts, etc.
Internat. bibliographies, descriptions of buildings, works of art, records in archives, etc.

**Figure 1: The national bibliography of the future can be modelled as a graph in the WWW.** *The ellipses are subgraphs. The arrows between the subgraps show possible interconnections between the nodes in the subgraphs.*

Thus, the national bibliography of the future cannot be a closed system and cannot without loss be isolated from the web. A printed version of the bibliography or even a complete backup of the national library's data would only be a lossy derivative, since the completeness of the national bibliography depends on data outside of the library's databases.

In this graph the nodes represent artefacts produced by an intellectual or artistic endeavour. Those artefacts – we will call them "cultural assets" – are essentially equivalent to the FRBR group 1 entities.[7] Each of those nodes bundles significant properties of the objects they represent and uses them to describe it. Those bundled properties have much in common with the descriptions we can find on cataloguing cards in that they are data (or facts) captured according to a particular set of rules. A publication's main title is still the same character string and its ISBN is the same sequence of numbers. Most other properties, however, are references to other entities, i.e. directed edges connecting one node to other nodes in the WWW. Those other nodes can be metadata nodes, too: descriptions of cultural assets, persons, corporate bodies, events, places, concepts, etc.

[7] For a description of the FRBR entities cf. IFLA (2009)

Several of those nodes are edited by libraries (e.g. GND,[8] RAMEAU,[9] and LCSH[10]). Possible are also references from library metadata to other nodes, e.g. links to the object itself (if a WWW resource) or references to nodes in datasets curated by organisations outside of the library community such as links from a name authority to the corresponding Wikipedia entry.

From this point of view, the German National Bibliography already is a heavily inter-linked dataset in a graph referred to as the data pool for the Deutsche Nationalbibliografie. When we use the term "data pool" we refer to the complete graph available today and in the future. I. e. the data pool consists not only of the data for which the Deutsche Nationalbibliothek is editorially responsible, but of all datasets related to the data of the Deutsche Nationalbibliografie. As the Deutsche Nationalbibliothek opens its datasets to the WWW (e.g. by offering linked open data services), our data pool will inevitably contain an increasing amount of non-library data. The relations in the data pool originate from several mechanisms: through the use of international standards numbers such as ISBN or ISMN there are implicit relations to other library items with the same standard number; relations created by mapping controlled vocabularies or authority files either intellectually (e.g. GND, LCSH, and RAMEAU) or by machine algorithms (e.g. VIAF[11]); or relations coming from co-operations with third party organisations (e.g. Wikipedia).

Boundaries and limits

It is difficult to define, where exactly a national bibliography starts and where it ends, but we can fairly easy find supersets of data containing all information relevant for its creation. At the top level a national bibliography is a proper subset of all descriptive data in the WWW. Further, it is a subset of all descriptive data in the cultural context. This data includes e. g. descriptions of persons, objects, concepts, and places that are relevant in a cultural heritage context. In addition, there are constraints such as the limitation to a particular geographic region (the nation) and the restriction to specific types of cultural heritage objects (e. g. books, maps, postcards, sound recordings etc.)[12].

[8] The Gemeinsame Normdatei (GND – Integrated Authority File) is a German authority file which covers all types of entities and which serves as a common, authoritative reference system for libraries' bibliographic data and for the cataloguing data of other authority file users such as archives, museums, projects, scientific and cultural institutions. See http://www.dnb.de/EN/Home/home_node.html

[9] RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) is a documentation language used by the Bibliothèque nationale de France, the French university libraries and several public libraries in France. See http://rameau.bnf.fr/

[10] Library of Congress Subject Headings (LCSH) are used to index materials at the Library of Congress and world-wide. See http://id.loc.gov/authorities/subjects

[11] VIAF – The Virtual International Authority File (http://viaf.org/) "is an international service designed to provide convenient access to the world's major name authority files." Cf. http://www.oclc.org/viaf/

[12] For a further discussion of this, cf. IFLA (2009)

Another aspect of the boundaries of a national bibliography is the administrative responsibility. In order to satisfy the requirements for long term reliability and high data quality, the responsible body must have sufficient governmental backing and be adequately equipped to manage and process the data for the foreseeable future. For a national library or national bibliographic agency, this is at the core of their *raison d'être* and should be part of their duties, often – but not always – supported by legal deposit legislation.[13] This does not infer – however – that all data curated by a national bibliographic agency is automatically part of the national bibliography: The Deutsche Nationalbibliothek manages several datasets, including special collections, which have no relevance for the German National Bibliography.

Contents

During a workshop at the Deutsche Nationalbibliothek, the participants should define what in their opinion constitutes the German National Bibliography. The answers ranged from "all data the library curates" to "only titles and excluding special collections". However, even the group arguing for the minimal solution was not prepared to abandon authority data completely. This is not surprising since the authority files are a product of the cataloguing process and contain information – e. g. the author name – that is necessary for a correct ISBD display, i. e. without this information the bibliographic record is not meaningful enough. On the other hand are not all data in the authority record necessary for the ISBD display of a bibliographic description in a printed national bibliography.

This leads to the assumption that it is not possible to answer the question, which data makes a national bibliography, on the entity level, but rather on the level of individual characteristics: Not all characteristics needed for the record of a cultural heritage entity are those of that entity itself, but rather of other entities attached to it.[14] The same applies to the search environment. Ideally an online catalogue should use all available data to optimise the search process. This can include data from outside of the library sphere that was created without any reference to bibliographic entities. One example is the use of geographic coordinates from datasets like geonames.org that can allow us to search by geographic locations or geographic proximity.

ISBD display and search/retrieval in an online catalogue are only two of many services surrounding the national bibliography. Not all data used in those two services are necessarily considered part of bibliography and thus we cannot derive what constitutes a national bibliography by looking at the services (printed bibliography, online search, national bibliographic services) we can build on top of it.

[13] Cf. Andersen (1974) p. 11.

[14] This is essentially an extension of the statement that a "bibliographic description typically should be based on the item as representative of the manifestation and may include attributes that pertain to the embodied work(s) and expression(s)" from IFLA (2009) p. 4.

When we build (national) bibliographic services on data that is not exclusively used for the national bibliography, the consequence must be to cut the dependency between service and data altogether. A clear and consistent picture emerges if we focus on the core of the national bibliography: publications. That way we arrive at a minimal data set of textual elements and references, which might not be very useful if taken on its own. Looking further towards FRBR[15] and RDA we can state this more precisely: a national bibliography records manifestations, i. e. "physical embodiment[s] of an expression of a work"[16]. The work, expression, and item data – as the authority records – are members of their own datasets, and the data in the national bibliography only contains references to those other entities.

### Processes for creation

National bibliographies increasingly should become part of a global ecosystem of data consumers and data producers. The stronger it is interlinked with domain independent data from a wider range of institutions, the more applications will profit from its knowledge base. The services libraries supply will increasingly build on data from external sources. The availability of additional information on entities related to cultural heritage – it be cultural assets, persons, or concepts – allows us to improve our bibliographic services. For parts of the data, it could even be possible that we refrain from curating the data ourselves and instead re-use third party information.

[15] Functional Requirements for Bibliographic Records. Cf. http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[16] IFLA (1998) p. 21.

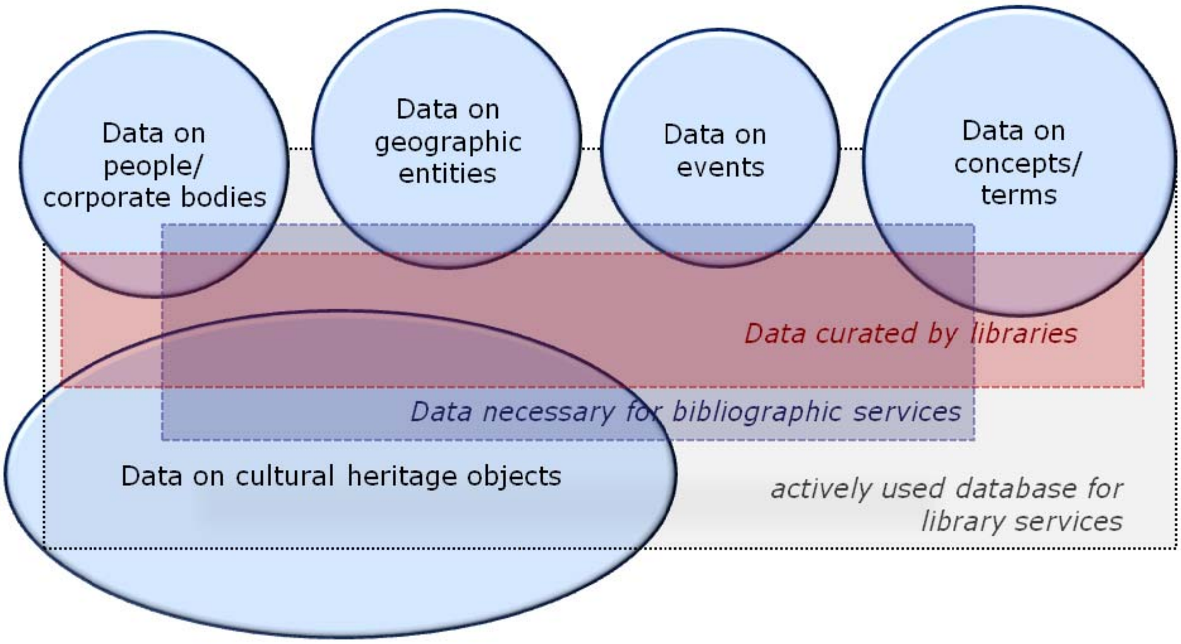
Data on people/ corporate bodies
Data on geographic entities
Data on events
Data on concepts/ terms
Data curated by libraries
Data necessary for bibliographic services
Data on cultural heritage objects
actively used database for library services

**Figure 2: Possible future distribution of data creation and data usage in libraries.**

1. Not all pieces of data libraries catalogue are necessary for their bibliographic services. Nonetheless they might be indispensable, e. g. for circulation purposes or archiving.
2. Not all pieces of data necessary for bibliographic services are produced/catalogues by libraries.
3. The data basis used by libraries is larger than the union of those two data sets.

This approach is not entirely new. At the Deutsche Nationalbibliothek we even today do not create all parts of the bibliographic description manually, but re-use third party information such as metadata supplied by publishers, or semi-automatically created information like tables of contents or machine-generated subject headings. With this background, the extension of the data re-use and a controlled opening of cataloguing to selected web communities seems a logical next step.

For libraries this is less about resigning on particular tasks and more about focusing on our own strengths, such as high quality data curation, knowledge organisation, and authority control. The result will be a more efficient division of work where each data provider can concentrate on their core competencies while sharing the data as a common good. The national bibliographic centre is responsible for published cultural heritage objects in textual form, for all references from those objects to other entities, and for their properties if they are necessary for the bibliographic services and if no other institution can provide them in an adequate quality. Other data providers are responsible for other properties.

Due to the increasing number of publications, libraries will have to restrict the manual creation, maintenance, and verification of metadata to those areas where those processes add value compared to automatic data creation. The definition of what is added value will in those cases not depend conformance to particular cataloguing rules, but will instead focus on increased usefulness for data selection and filtering on the one hand and the library customers' business models on the other hand.

## Quality characteristics

In order to ensure that the quality of library services does not suffer from the integration of third party data, but – on the contrary – can profit from it, libraries need to consider carefully which external data sources they re-use. The reliability (referentiability, persistency, use of standards, and quality assurance) of the data is crucial. The origin of the data and – if available – the process behind its creation must be transparent, and there has to be quality control processes in place. Particularly each external dataset must be evaluated regarding its long-term perspective.

The re-use of third party datasets in a national bibliography can cause a conflict with existing cataloguing rules. Many – or even most – data suppliers outside of the library community are not aware of AACR II, RAK, or RDA, and we cannot expect them to invest even minimally in that area. Instead we need to purge the existing cataloguing rules, and evaluate them with the question if the rule helps the national bibliography fulfil its intention: to serve as a comprehensive record of a country's publication landscape. Only if the rule does, we should use it as a benchmark when evaluating the quality of third party datasets.

When an institution uses third party metadata it needs to define what constitutes a trustworthy data provider. In order to do this, a list of criteria similar to those that exist for e. g. trustworthy digital archives is necessary. When generating this list, the main criterion should be if the data provided matches the requirements for interoperability. Data not deriving from trustworthy providers need to be marked accordingly.

The same applies to machine-generated data. Here, it is important that the processes are well documented and that the results are reproducible. If we can produce homogenous data for subsets of the national bibliography, the consumer acceptance will grow enormously. One way of achieving this could be the collaborative compilation of a list of criteria for trustworthy automatic processes. It comes without saying that it will be necessary to monitor those processes regularly.

## Technical realisation

Above, we discussed that a national bibliography can be built on a minimal set of manifestation metadata combined with information from authority files and interlinked datasets. This, together with the proposition that the bibliography is a graph in the World Wide Web, makes it a natural consequence to use linked data techniques – essentially an http-based framework for machine-readable relations between entities – for the technical realisation of the bibliography.

Looking at Tim Berners-Lee's requirements for linked data,[17] the implementation is fairly straight-forward:

1) Use URIs as names for things
   Each entity we use in the national bibliography needs its own unique identifier. This applies at least to manifestations, controlled vocabularies, corporate bodies, and persons; ideally also to other authority data such as publishers.
2) Use HTTP URIs so that people can look up those names
   When libraries assign identifiers to their entities, they should leverage the infrastructure of the WWW.
3) When someone looks up a URI, provide useful information
   In order to maximise re-use of library data, it is imperative that we provide concise descriptions of the data node identified by the URIs when someone dereferences that identifier. By doing this we invite other data providers to re-use library datasets by linking to authority data or bibliographic information using our URIs.
4) Include links to other URIs so that they can discover more things
   When we re-use authority data or data from other content providers in our bibliographic records, we may choose to display *some* properties of the linked entity, e. g. an author name or the preferred form of a subject heading, together with the manifestation's title, while still offering a link to a page where the user can find more information, such as date of birth, or profession.

When building this infrastructure it is not important if the data modelling is done using RDF or any other technical means, linked data per se being format agnostic. The RDF stack of technologies does have a huge potential to make data and semantic connections machine-readable (and thus machine-understandable) in a generic way. This will, however, in many cases lead to a highly complex data model, and the problems that can occur when using RDF for deployed applications have been cause for controversial discussions.[18] At the Deutsche Nationalbibliothek, we have also had the experience that there are use-cases, e. g. temporal relations or graph provenance, where more expressivity is needed than RDF can deliver without using overly complex models – at least at the moment. This might change as RDF modelling matures and concepts like named graphs become more prevalent. As long as we consistently build our data models on the linked data principles – publishing our data in the WWW under stable identifiers and enriching those representations with references to other entities – we pave ourselves the way for the introduction of more expressive description languages.

[17] Berners-Lee (2006)

[18] As an example for many, see Miličić (2011) and the discussion there.

## Usage

When searching for information on the web, most users turn to commercial search engines like bing, ask.com, and Google, or to social networks like LinkedIn, and Facebook.[19] For (national) libraries and other cultural heritage institutions, this means that we must give those companies the possibility to creatively integrate our data into their services so that users can profit from this information during their web search. For example the European cultural heritage portal Europeana experienced that creating a specific landing page for each object in the collection and giving that page a unique URI improved the visibility to and increased the traffic coming from search engines.[20]

This shows that the library's own online catalogue will continue to play an important role in the information landscape. As a highly specialised search engine it is the benchmark application showing what can be done with library data and how that data can be tailored for specific user groups. On the other hand, the border between the added value offered by a library catalogue and a specialised web search engine becomes increasingly blurred. We can imagine that in future some search engine providers will focus on bibliographic information, e. g. for specific science communities. The availability of high quality bibliographic data embedded into the environment where the users search for information will be beneficial to our patrons and will make it easier for them to find, select, and obtain the appropriate resources.

In order for the library data to be easily re-used in other services it is necessary to explicitly state the licensing terms. Particularly for a national bibliography, added value is created through increased visibility and re-use of the data, and the easiest way of encouraging this is to offer the data under a flexible and open license that explicitly allows commercial re-use. This can be controversial, considering that large players like Google financially are in a position to pay for library metadata. It is – however – questionable if they would, and if the library information is not visible in the major search engines, it will be invisible to most people searching the web. The use of an open license also allows the integration into open platforms such as the Wikipedia and small and medium-sized websites such as online news portals. That way, the increased visibility will definitely outweigh the (potential) loss of income.

[19] According to a recent PEW report, 92% of online adults use search engines to find information. Cf. Purcell (2011)

[20] Clark et. al. (2011) p. 15-17

## Current work at the German National Library and implications for the German national bibliography

The German National Library is – as other national libraries, too – working on the implementation of the necessary changes.

Starting in 2008, the Deutsche Nationalbibliothek is successively publishing its entire database as linked open data; currently circa 70% of the title data is available. The first dataset was the authority data from the GND, followed by the German translation of the Dewey Decimal Classification. After a six month project, we were able to distribute the major parts of the title data from the national bibliography as linked data. Except for the DDC data, which is available under a CC BY-NC-ND license, all other data from the Deutsche Nationalbibliothek is published using CreativeCommons Zero.

Further, we increasingly use automatic processes to create descriptive and subject metadata. One line is to enrich the bibliographic descriptions in the catalogue e. g. with tables-of-contents, and we intensify the inclusion of metadata from publishers and universities, including subject categories – e. g. BISAC codes – and subject headings. For online publications in the national bibliography's series O, we completely re-use the descriptive and subject metadata supplied by the content depositors.[21]

There is a current project where we evaluate the possibility to automatically add subject headings to bibliographic records. The process is built on top of an automatic indexer. We are analyzing how it handles different document types and the quality of search for automatically indexed publications. First results will be available in 2013.

Automatic indexing of web resources will be done with the same controlled authority headings as we use for intellectual indexing procedures to keep searching for information as homogenous as possible. Thus, reliable authoritative data can be offered. Furthermore, linked information of people, places etc., but also semantic relationships within the authority data can be used efficiently.

Today's users have diverse expectations of bibliographic data especially of subject access, which includes an overview of available literature, bibliographic citations, or direct access to publications. At the Deutsche Nationalbibliothek, we follow the IFLA conviction that a coherent indexing policy and the use of controlled access remain important in providing order and consistency of data.[22] All our users can benefit from having well-organized subject structures - independent from data formats, distribution channels or representation displays in which they use the national bibliography. Classification systems and subject headings support users in reaching (find, identify, select or explore) the information they want. A majority of users (end-users or professional ones) is

[21] For further information on the handling of online publications in the Deutsche Nationalbibliothek cf. Gömpel and Svensson (2011)

[22] IFLA (2012)

interested in partial data sets rather than the entire bibliography. Most users may not even know that they are searching in the bibliography when searching within our online catalogue. An increasing number benefits from personal profiles in searching the bibliography or subscribe RSS-feeds of query results. Additional linked content information as for instance tables of contents or abstracts help users to understand the information they searched for. The provision of linked online content, especially from e-books and e-journals is appropriate to users' needs.

Due to the growing number of published information it becomes more important to categorize the information into manageable units that are readable, selectable, and can be searched precisely. Ideally, complete and detailed indexing is applied to all documents to help users find relevant search topics or additional content information. Faced with the enormous amount of printed publications in Germany and the growing number of published web resources under the .de-domain but also of budget and staff restrictions, we decided on a gradual indexing policy which enables us to manage the entire national output with different levels of detail. This modular policy[23] allows for a variety of access points, selecting different media types, but is transparent and controlled by quality criteria.

The minimal indexing level is a very broad classification number that is provided for nearly all resources – *Sachgruppe* / subject group which is equivalent to a rough classification number. The structure is based for the most parts on the two top hierarchical levels of the DDC (the Hundred Divisions or Second Summary). Some differences are made by integrating deeper levels to meet the local users' needs. Fiction and children's books can be selected separately as well as school text books.

Several other projects help to work on improving the linking to other datasets. An excellent starting point is the optimisation of our own data. It became evident, that the conversion from textual descriptions to references is very suitable for automation. Other processes for named-entity-recognition and the transformation of entity names to references are currently in an experimental stage. When creating links to external datasets, the Deutsche Nationalbibliothek deploys a mixture of co-operative, manual, and automatic processes. Through VIAF we could automatically connect persons from the GND to other authority files, whereas MACS[24] used an editorial process to link subject headings from GND, LCSH, and RAMEAU. A co-operation with the German Wikipedia constantly delivers links between Wikipedia articles and persons from the GND, and within

[23] For an overview of the different distribution channels for bibliographic data in the Deutsche Nationalbibliothek see Svensson and Jahns (2010)

[24] Multilingual Access to Subjects (MACS) is a project with the objective to develop a system that allows multilingual subject access to library catalogues using existing indexing languages. It currently uses RAMEAU, LCSH, and the topical subject headings from the GND. Cf. http://www.nb.admin.ch/nb_professionnel/projektarbeit/00729/00733/index.html?lang=en

CONTENTUS we develop a technique to build a network of geographic names from the GND, Wikipedia articles, and geonames.org.

In order to drive this development further, the DNB has together with hbz, the Northrhine-Westfalian Library Network, founded the platform culturegraph.org. The underlying idea is to provide a technical and organisational infrastructure to support the interlinking of bibliographic data, thesauri, classifications, and other authority data. The resulting information network shall also be published as linked data. A core attribute of the platform is the support for citeability. Currently, the platform contains bibliographic descriptions for all publications since 1945 from the German library networks. A first intermediate goal is to create data clusters to identify which bibliographic records describe the same manifestation and to name and publish those clusters using a unique, common identifier. Further evaluation will show, if it is possible to create clusters for works (in the FRBR sense) and if that data can serve as a seed for an authority file for intellectual works.

All those national activities need to be integrated into an international context, and the possibly most important task will be the continued participation in international projects and standards bodies working on the future of cataloguing and data exchange.

## Conclusions and future work

In this paper, we argue for a library data ecosystem built on linked data principles, published in the World Wide Web under an open license. In order to spread this data as widely as possible, we should aim to co-operate with major search engine providers and have them integrate bibliographic data into their databases.

In order to serve a wide variety of customer needs – both in terms of data selections and formats – it is necessary for the underlying database to be as flexible as possible. To use an analogy from indexing, we need to move from pre-coordinated data to an information architecture in which we can post-coordinate bits and pieces of information according to ever-changing requirements.

Current discussions regarding cataloguing rules must focus on the needs emerging through the inclusion of bibliographic data in the WWW. RDA has as its goal to be "a new standard for resource description and access designed for the digital world".[25] Despite rightful criticism that it is still too heavily biased towards traditional cataloguing categories and does not push far enough forward, it is a step in the right direction given the increased flexibility it offers. There is no doubt, however, that further revisions are needed.

Perhaps even more important than library cataloguing rules is the possibility to interact with data from outside of the library ecosystem. Much available and

[25] Cf. the strategic plan for RDA 2005-2009. Online available at http://rda-jsc.org/stratplan.html

useful data – e. g. a publisher's advertisement for a new publication that might contain words relevant for a user search – is neglected for search and retrieval purposes. This requires that libraries have means to store the provenance of the third party data and to make this information transparent to end users.

Further it is necessary to make citeability a quality criterion. If a bibliographic description is marked as "citeable" it must never be deleted and it will be possible to trace all changes ever made to it. In order to enable this, we need to enhance the processes for data creation and maintenance and add provenance information for all data emerging from automated processes.

Finally we must invest more in linking library data to other datasets. Wherever possible, we must use references to authority data instead of textual content ("literals"). An international co-operation for common authority files – e. g. VIAF – will be a major step in the right direction and will help avoiding duplicate work. A central repository of work titles will help tremendously with connecting bibliographic descriptions across national borders and languages. A database of publishers (a subset of corporate bodies) does not only provide controlled access points, but will also promote cooperation between libraries and publishers.

The national bibliography should be a graph in the web of graphs. The digital age has caused the information sector to change. Instead of hanging on to our existing structures and practices, libraries should embrace the change and instead try to act as change agents. More flexibility and openness are a good start.

## References

Anderson (1974) Anderson, Dorothy: Universal bibliographic control : a long term policy; a plan for action. Pullach 1974.

Berners-Lee (2006) Berners-Lee, Tim: Linked Data. Available at http://www.w3.org/DesignIssues/LinkedData.html

Clark et. al (2011) Clark, D. J.; Nicholas, D.; Rowlands, I.: D3.1.3 – Publishable report on best practice and how users are using the Europeana service. Available at http://www.europeanaconnect.eu/documents/D3.1.3_eConnect_LogAnalysisReport_v1.0.pdf

Gömpel and Svensson (2011) Gömpel, Renate; Svensson, Lars G.: Managing Legal Deposit for Online Publications in Germany. 2011. urn:nbn:de:101-2011061609. Online available at http://nbn-resolving.de/urn:nbn:de:101-2011061609

Heath and Bizer (2011) Heath, Tom; Bizer, Christian: Linked Data: Evolving the Web into a Global Data Space. San Rafael (Calif.) 2011. (Also online at http://linkeddatabook.com/editions/1.0/)

IFLA (1998) Functional requirements for bibliographic records : final report. München 1998. All references in this paper are to the 2009 revised online version at http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

IFLA (2009) IFLA Cataloguing Principles: Statement of International Cataloguing Principles (ICP) and its Glossary. Ed. Barbara Tillett and Ana Lupe Christián. München 2009. For a list of online versions (incl. translations) see http://www.ifla.org/publications/ifla-series-on-bibliographic-control-37

IFLA (2012) , Guidelines for Subject Access in National Bibliographies. Ed. Yvonne Jahns. Berlin: 2012.

Miličić (2011) Miličić, Vuk: The Ultimate Problem of RDF and the Semantic Web. Blog post, available at http://milicicvuk.com/blog/2011/07/19/ultimate-problem-of-rdf-and-semantic-web/

Purcell (2011) Purcell, Kristen: Search and email still top the list of most popular online activities: Two activities nearly universal among adult internet users. 2011. Available online at http://pewinternet.org/~/media//Files/Reports/2011/PIP_Search-and-Email.pdf

Schuster and Rappold (2006) Schuster, Michael; Rappold, Dieter: Social Semantic Software – was soziale Dynamik im Semantic Web auslöst. In: Semantic Web: Wege zur vernetzten Wissensgesellschaft. Ed. Tassilo Peregrini. Berlin 2006.

Svensson and Jahns (2010) Svensson, Lars G.; Jahns, Yvonne: PDF, CSV, RSS and other acronyms: redefining the bibliographic services in the German National Library. 2010. Available online at http://www.ifla.org/files/hq/papers/ifla76/91-svensson-en.pdf

urn:nbn:de:101-2012052306