IN2N: Cross-institutional Authority Collaboration

Alexander Haffner
German National Library,
Germany
a.haffner@dnb.de

Abstract

The paper describes the work being conducted in the Cross-institutional Authority Collaboration (Institutionenübergreifende Integration von Normdaten, IN2N) project. This pilot project, executed in cooperation with the German National Library and the German Film Institute, aims to establish new collaboration models to improve cross-domain authority maintenance. The paper outlines applied strategies for providing a shared infrastructure as well as workflows for exchanging data about persons; interface enhancements permitting the exploitation of innovative web approaches; and cross-institutional data search and representation solutions. Furthermore, we discuss specific boundary conditions, such as disparities in the level of data granularity, for an interoperable cataloguing environment.

Keywords: collaboration; authority data; cross-domain; match&merge; gnd; linked data

1. Introduction

Maintaining and linking authority data are essential components of descriptive and subject cataloguing. Over the last 15 years, German libraries have generated a huge amount of authority entries describing persons, corporate bodies, conferences and events, places or geographic names, topics, and works. The more than ten million authority entries held in the Integrated Authority File (Gemeinsame Normdatei, GND) are managed cooperatively by all library associations in the German-speaking countries. GND data is highly cross-linked on a national and international level to other library-specific authority files (Behrens-Neumann, 2012). Key examples are the Virtual International Authority File (VIAF) as well as the authority data of the Library of Congress. Additionally, the German Wikipedia community has aligned almost a quarter million articles to GND entries.

Authoritative control plays an important role in other domains beyond libraries. Today, the efficient information management of big data is almost impossible without structured and well organized datasets. Authority data from libraries can support the organization of data from other major players. In order to this, the following questions arise from the library perspective:

- Are there stakeholders that do the same work as libraries, and maybe even better?
- How can the work be shared?
- What collaboration models have to be established for partners from new domains to be able to participate in the authority maintenance process of libraries?
- Are we already fulfilling all the technical and organizational requirements for successful collaboration?

The German National Library identified the fields of cinema and film archiving as an interesting non-library supplement for integration into the GND. Data from intersections as well as from new entries promise sustainable enrichment for all GND participants. Similarly, the German Film Institute (Deutsches Filminstitut, DIF) has discovered the advantages of cooperative data maintenance for their process optimization.

Cross-institutional Authority Collaboration (Institutionenübergreifende Integration von Normdaten, IN2N) is a collaborative research project of the German National Library and the

German Film Institute. It is financially supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). The main goal is the initial alignment and linking of the existing authority entries for persons in both institutions' data and the establishment of an organizational and technical web-based infrastructure for continuous bi-directional data exchange. The cooperative data maintenance system will be based on independent and differentiating storage systems, data formats and data models. The experience gained in the project will be used to develop a generalized collaboration model for working with further non-library cooperation partners.

2. Initial Match&Merge Environment

In www.filmportal.de the German Film Institute has created a central, cost-free internet platform for providing reliable, in-depth information on all German cinema films. Besides offering an overview of the vibrant German film culture, from its beginnings up to the present day, filmportal.de provides information about the people in front of and behind the camera. Currently, the dataset comprises more than 180,000 differentiated persons.

The GND constitutes a conclusive reference system for bibliographic library data (references for print, multi-media and electronic resources) and for the cataloguing data of other authority data users such as cultural institutions. At the beginning of 2013, the GND comprised 2.8 million differentiated persons. Since July 2012 all authority data in the GND have been provided free of charge for use under the Creative Commons Zero license.

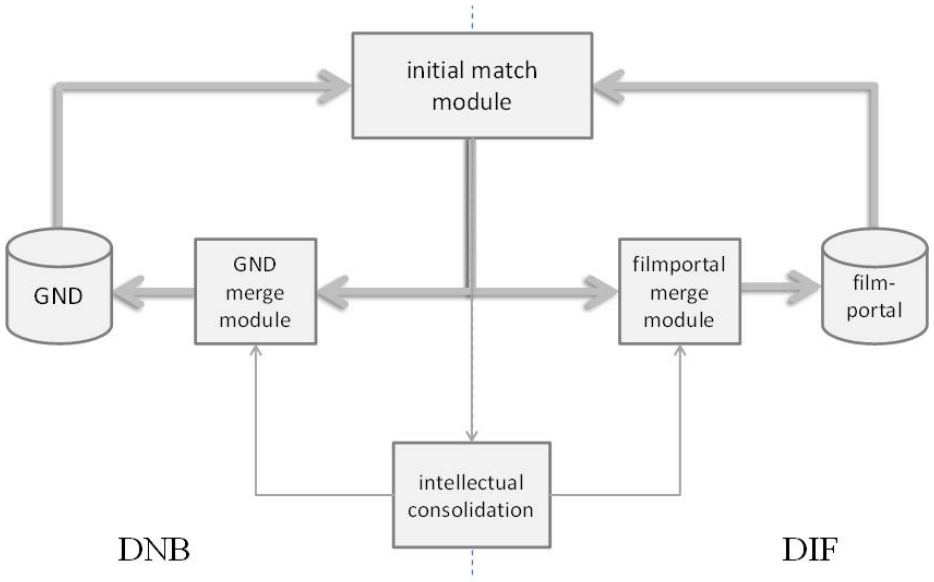

FIG. 1 Match&Merge Workflow for GND and filmportal data

Figure 1 illustrates the underlying workflow for processing both datasets in order to identify potential equivalences. The match process compares one person from filmportal.de against the whole GND dataset and calculates a match probability. The possible results can be divided into:

1. Exact equivalence between two persons,
2. Potential equivalences between persons, or
3. No equivalence to the corresponding dataset.

If an exact match is identified, both merge modules adopt information missing in their own data stores from the corresponding partner. If no match can be identified, a new entry is created in the GND for the person from filmportal.de. Subsequently, all personal data from filmportal.de

will be available to all GND partners. The most innovative part lies in the intellectual equivalence consolidation. Ambiguous matches have to be checked by the commissioning editor, and equivalence pairs chosen manually. The commissioning editor is always provided by the institution which intends to ingest its data into the GND. This organizational aspect is mainly due to human resource issues resulting from the expected increase of parallel initial Match&Merge processes.

It is to be emphasized that the partners have different needs with regard to the data ingest. These are mainly attributable to deviations in the cataloging rules, controlled values or the data model. As a consequence we create institutional responsibility in the subsequent use of the data as a desirable part of the aspired collaboration model.

The implementation is based on the components provided by the Culturegraph Metafacture Framework. The framework extracts highly efficient core characteristics needed by the match algorithm. Currently the match characteristics comprise names of persons (including all parts, titles and academic degrees), dates of birth and death, places of birth and death, and occupations. Unfortunately, we cannot include titles of films since this information is not yet stored in the GND. Besides the match characteristics we use a number of additional data elements for the data import; these include affiliation, period of activity (if no dates of birth or death are available), biographical or historical information, external identifiers, and of course information about films a person has been involved in. These elements are usually very divergent in two data sets, which makes incorporation in the match process almost impossible.

A certain challenge was posed by the comparison of names. The German Film Institute and the GND use different composition models of name parts. For efficient comparison we decided to generate sort names first that indicate the same name part order; if the sort names do not match, we append equivalent checks based on the single name parts.

Another special feature concerns the assignment of institution-specific terms for occupations. The German Film Institute applies literal descriptions from a list of controlled terms. The GND persons are linked to subject headings that do not fully align with the available film occupations. However, the terms of the common subset had to be assigned manually to perform accurate matching and to increase the data quality in the merge and data import processes. The situation is similar for geographic name assignments. Unfortunately, filmportal.de does not use a controlled list for this information, which makes it even more difficult to assign the corresponding GND records for geographic names or places. First assignments could be made by identified equivalents in GND and filmportal where both persons' places of birth or death are given. Also the resulting list of places needed manual double-checks.

We will undertake an evaluation of the match process, focusing on optimization options that are generated by enriching the GND dataset with third party information. The chosen candidates are VIAF and Wikipedia due to the fact that these already provide references to the GND. Efficient processing of these large datasets is also supported by the application of Culturegraph's Metafacture Framework.

In a first trial the equivalence check of persons from filmportal to persons described in German Wikipedia articles immediately discovered more than 10,000 matches. In contrast to GND, Wikipedia articles usually contain a person's filmographic information which enabled a comparison based on names of persons and movies they were involved in.

3. Cataloging Routine via the Web

3.1. Exchange Format

In current practice, all GND participants have to harvest and deliver data in GND/MARC format. This is basically an adaptation of MARC 21 for Authority Data that supports more efficient linking among authority entries. The same advantages are offered by the GND/RDF format, which provides the corresponding semantic web representation. In contrast to MARC, the

data elements are intuitively understandable due to self-explanatory element names. This approach minimizes barriers to re-use by non-library organizations. Furthermore, the GND Ontology[1], as the underlying specification, offers vocabulary alignments to the FOAF vocabulary as well as to the RDA Vocabularies, which helps boost interoperability (Haffner, 2012).

Originally, the project intended to use Encoded Archival Context – Corporate bodies, Persons, and Families (EAC-CPF) as a domain-specific authority format for archives. An analysis of current implementations has revealed that the standard provides increased scope for differences in the actual application, which causes interoperability issues among the datasets. The data providers concerned offer EAC-CPF data in locally enriched dialects. Especially the dialect of the German Film Institute was divergent to other major players, which makes reuse of the eventual infrastructure difficult without provider-specific enhancements. Accordingly, EAC-CPF cannot be considered as a sustainable format for the IN2N project. The goal of establishing a generic collaboration model cannot be achieved based on a heterogeneous format.

On the basis of the results obtained, we strongly believe that GND/RDF is the format which should be applied. In practice this means that partners can retrieve information in GND/RDF and provide GND/RDF for update operations.

We have also examined update approaches that avoid the use of established exchange formats like MARC 21, EAC-CPF, or even GND/RDF. Non-library organizations usually wish to create rudimentary new data entries or add or change single facts in existing data entries. The idea is that partners should only send differences to the current GND entry in a very simple expression. The approach has been adopted from revision control systems.

3.2. Web Interface

In contrast to today's data exchange practices, IN2N avoids OAI harvesting principles. Mirroring millions of data records while maintaining only a small percentage of them would not seem to be an adequate approach for attracting new partners in non-library domains.

We aim to follow the REST paradigm by supporting GET, POST, PUT and DELETE operations via URL based requests. Therefore we have built on Search/Retrieval via URL (SRU) techniques. SRU is a standard search protocol for search queries via URL which utilizes the Contextual Query Language (CQL) for the compilation of corresponding results.

Using current update mechanisms, editorial systems fetch a complete record, edit it, and import the record back into the source system. For the MARC 21 format it is an appropriate approach, whereas for RDF-based formats it means that only a subset of the original descriptive information will be available and all administrative information from the record will be lost.

As a consequence, we are also researching into alternative web-based update strategies to ingest data on a highly granular level. These updates take place at the property level and, as mentioned before, only comprise differences to the currently stored data entry. We have limited the operations offered to add, change, and delete. Change and delete operations identify the relevant field in the internal system by committed values. For instance, if the spelling of a name has to be changed, the change operation has to identify the misspelled name by listing all name parts and their values as well as the new spell-checked name with all its parts. To create a new data entry the request only lists a number of add operations without referring to a specific GND data record.

By providing simple and efficient search as well as update interfaces we expect to lower the threshold for the implementation of non-library editorial systems accessing the GND as their authority reference system.

[1] http://d-nb.info/standards/elementset/gnd#

4. Cross-Dataset Search

Authorities are also playing an increasingly important role in the field of semantic search, as they provide reliable access points and identifiers for entities. Culturegraph.org is analyzing major bibliographic catalogs from Europe as well as crosslinking data from various sources to make equivalences and relationships available. Furthermore, Culturegraph.org acts as a datahub: it provides a cross-domain and cross-dataset entry point for searching and browsing based on authority files - thus also increasing the visibility of all participating institutions. Improvement of the visibility and accessibility of the network is the next major milestone in supporting the implementation of integrable exploration and visualization applications.

IN2N will support the platform in order to benefit from its developments. IN2N will provide authority data, bibliographic and filmographic data to Culturegraph.org. After data ingest, search requests can be sent to Culturegraph's REST interface and result sets will be dynamically integrated into the local catalog's result representation. IN2N aims to investigate how helpful this approach is to end users. With regard to usability, our intention is to find the right balance between local and remote information and increasing the user's search success.

5. Conclusion and Outline

In late 2013 we intend to effect the initial matching and integrate the merge modules into our workflows. In 2014, the German Film Institute will migrate to its new editorial system which accesses the GND through the newly established web interfaces. We will also finalize all merge operations, including the intellectual consolidation, at this point. In the last quarter of the project, from May to November 2014, we will devote our energies to acquiring additional partners on the basis of the results achieved.

During the course of our work so far we have come to recognize that such a project can work on independent and different database systems and internal data formats, however working with inconsistent data models can cause substantial problems. Customization of the filmportal.de data model was necessary to make it compatible with the highly granular GND data in order to allow match and search operations. For example, parts of names had to be consolidated as well as entity types. Otherwise data imports would cause frustration rather than increase efficiency in daily operation.

Library rules are sometimes a blessing, sometimes a curse. However, granular data simplify matters in the challenges we face in the IN2N project.

References

Behrens-Neumann, Renate. (2012). Die Gemeinsame Normdatei (GND) – Ein Projekt kommt zum Abschluss. Dialog mit Bibliotheken 2012/1. Retrieved, March 25, 2013, from http://nbn-resolving.de/urn:nbn:de:101-2012100846.

Haffner, Alexander. (2012, July 16). Internationalisierung der GND durch das Semantic Web. Project Report of the Competence Center for Interoperable Metadata. Retrieved, March 25, 2013, from http://www.kim-forum.org/Subsites/kim/SharedDocs/Downloads/DE/Berichte/internationalisierungDerGndDurchDasSemanticWeb.pdf.