IN2N: Cross-institutional Authority Collaboration

Alexander Haffner (DNB)
The IN2N Project

- research project, executed in cooperation with:
  - the German National Library and
  - the German Film Institute

- duration
  - December 2012 - December 2014

- financially supported by the German Research Foundation
Authority Collaboration in the German-speaking Library Community

- Collaborative maintaining and linking authority data are essential components of descriptive and subject cataloguing.

- Integrated Authority File (Gemeinsame Normdatei, GND)
  - More than 10 million authority entries.
  - Describing persons, corporate bodies, conferences and events, places or geographic names, topics, and works.
  - Aligned to VIAF, German Wikipedia etc.
  - Accessible as Linked Open Data.

- BUT
  - Data exchange based on harvesting strategies.
  - Data model and data formats are very library specific.
  - Non-library organizations are almost excluded.
Cross-institutional Authority Collaboration

- assumption: authority data from libraries can support the organization of data from other major players

- arising questions from the library perspective:
  1. Are there stakeholders that do the same work as libraries, and maybe even better?
  2. How can the work be shared?
  3. What collaboration models have to be established for partners from new domains to be able to participate in the authority maintenance process of libraries?
  4. Are we already fulfilling all the technical and organizational requirements for successful collaboration?
IN2N Objectives

- initial alignment and linking of the existing authority entries for persons in filmportal.de (180,000) and GND (2.9 million)

- establishment of an organizational and technical web-based infrastructure for data exchange based on differentiating
  - storage systems,
  - data formats, and
  - data models

- development of a generalized collaboration model for working with further non-library cooperation partners

- use of Linked Open Data
Major Phases for an Active Collaboration

1. Initial data match between the partners' data set and the GND data, and a succeeding bi-directional data import from information missed in the respective data stock.

2. Cataloging routine via a web interface to perform GND queries in real-time and to update GND entries by transmitting differences to the currently stored data entry.
Phase 1: Initial Match&Merge

- Initial Match Module
- GND Merge Module
- filmportal Merge Module
- Intellectual Consolidation

IN2N: Cross-institutional Authority Collaboration | DC-2013 | 4. September 2013
Phase 1: Initial Match

- match characteristics
  - entityType, name, dateOfBirth, dateOfDeath, gender, placeOfBirth, placeOfDeath, occupation

- match results can be divided into:
  1. Exact equivalence between two persons,
  2. Potential equivalences between persons, or
  3. No equivalence to the corresponding dataset.

![Venn diagram]

1. 2
2. 3
identify criteria for class membership
Phase 1: Improvement of the Match Process

- match process evaluation
  - iterative configuration of the match algorithm
  - enriching the GND dataset with third party information

- German Wikipedia as discovery aid
  - comprises 250,000 GND references
  - provides person's filmographic information
  - executed an equivalence check between filmportal.de information and Wikipedia’s person templates as well as article texts
    - discovered more than 10,000 GND matches

- Culturegraph’s Metafacture Framework
  - powerful tool suite for metadata processing
Phase 1: Intellectual Consolidation

- easy to handle user interface to make quick equivalence decisions
- Web UI with person’s main characteristics and match score assignments
- links for further research
  - i.e. filmportal.de, GND, Wikipedia, VIAF
- re-use of EntityFacts
Phase 1: Initial Data Import

- merge characteristics complementing the match characteristics
  - titleOfNobility, academicDegree, periodOfActivity, affiliation, geographicAreaCode, biographicalOrHistoricalInformation, homepage, contributedWork, externalIdentifier

- partners have different needs with regard to the data ingest
  - i.e. deviations in cataloging rules, controlled values
  - consequence: institutional responsibility for data ingest

- DNB as responsible party for the GND has to define access restrictions
  - it is allowed to overwrite a date of birth but not to delete an existing one
  - if place of birth contains a link to a geographic entry it is not allowed to replace it by a literal information
  - …
Phase 2: Cataloging Routine via the Web

- **goal**
  - lower the threshold for the implementation of non-library editorial systems accessing the GND as their authority reference system
  - providing a simple and efficient GND search as well as update interface

- **data format**
  - limited data set (approx. 25 elements for persons)
  - self-explanatory element names

- **property-based search functionality on GND data**
  - use of widely applied standards

- **updates without knowledge of the complete corresponding GND record**
  - incremental approach vs. record based approach
Phase 2: Use Case

- Local person search by editor without success
- Remote search via the person's name
- Result set transmission
- Editor selects entry from result set
- Partial data ingest into local database
- Data adaptation by the editor
- Transmission of changes
Phase 2: Applicable Data Formats

- EAC-CPF/XML
- GND-MARC-Format
- GND/RDF
- RDA Vocabularies
- BIBFRAME for Authorities
Phase 2: Update Interface on Property Level

- REST-Interface with a JSON transmission format
- operations
  - add, change, and delete

PUT uri="http://d-nb.info/gnd/129952788"
  add name(name.forename="Wolke A."); name.surname="Hegenbarth")
  add dateOfBirth(dateOfBirth.year="1980"); dateOfBirth.month="05");
  dateOfBirth.day="06")
  add placeOfBirth="Meerbusch, Deutschland"

PUT uri="http://d-nb.info/gnd/129952788"
  change
  name(name.forename="Wolke A."); name.surname="Hegenbarth")
  name(name.forename="Wolke Alma"); name.surname="Hegenbarth")
  change
  placeOfBirth="Meerbusch, Deutschland"
  placeOfBirthUri="http://d-nb.info/gnd/2029013-5"
Cross-Dataset Search

- **Culturegraph.org**
  - acts as datahub for searching and browsing
  - analyzes major bibliographic catalogs and crosslinking data to make equivalences and relationships available
  - authority data as access points

- **IN2N will support the platform in order to benefit from its developments**
  - providing authority data, bibliographic and filmographic data
  - re-use of Culturegraph’s REST interface for search
  - dynamically result integration into local catalog’s representation

- **usability evaluation**
  - find the right balance between local and remote information
  - increasing the user’s search success
Timeline

- **2013**
  - implementation of match environment
  - implementation of property-based update interface
  - GND Change Notifier

- **1st Quarter 2014**
  - Web-UI for intellectual consolidation
  - initial startup of the extended filmportal.de editorial system
  - RDF representation for entries from filmportal.de

- **3rd Quarter 2014**
  - cross-dataset search
  - acquiring additional partners
Conclusion

- Cross-institutional collaboration on independent and different database systems and data formats is possible.

- Inconsistent data models can cause substantial problems:
  - Customization sometimes is necessary.

- Highly granular data supports the collaboration.

- Decisions to be made by libraries:
  - How valuable is “library rules compliant data”?
  - Are we prepared to compromise?
  - Are we already open-minded enough to let external partners touch our data?
Vielen Dank!

discussion is welcome…

www.in2n.de

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