From raw data to rich(er) data

Lessons learned while aggregating metadata

Julia Beck | j.beck@ub.uni-frankfurt.de | @j4lib

SWIB 2019
Session: Aggregation and Interlinking

26.11.2019
Back to 2016 – What this talk will be about

• Review 2016

• What worked out and what did not?
• Which challenges did we face then and which do we face now?
• What does the metadata management workflow look like today?
• Not every challenge is solved yet,
  so we are looking forward to feedback and suggestions for tools
Specialized Information Service Performing Arts

„Past forward“
Project documentation
Recording, 2018
[Tanzfonds Erbe]
Specialized Information Service
Performing Arts

• Aggregates metadata from GLAM institutions from the performing arts domain (at the moment especially German-speaking institutions from Germany, Austria and Switzerland)
• Funded by the German Research Foundation
• What we are doing is best seen here:

• And here:

http://www.performing-arts.eu
Specialized Information Service
Performing Arts

search find discover

Behalte die Filtereinstellungen.

Suche:

Suche einschränken

Nur in Archivbeständen suchen

Treffern 1 - 20 von 12.699 für Suche „An sämtliche geehrtesten Herren Musik- und Theater-Freunde, Lieben und Musiker, zum neuen Jahre 1818“, Suchdauer: 0,10s

Sortieren

Relevanz

1. Fridolin, oder Der Gang nach dem Eisenhammer : mit obrigkeitlicher Bewilligung wird heute Mitwoch den 15ten April 1818 aufgeführt ; Schauspiel in fünf Aufzügen
   In Theater in Düsseldorf
   Holbein, Franz Ignaz von (Beitragende), Dragheim, Karl (SchauspielerIn), Trost, 1818
   1 Plakat : sw ; 33 x 19 cm
   Druckschrift ; Theaterzettel
   Universitäts- und Landesbibliothek Düsseldorf

2. Die Entführung aus dem Serail oder Belmonte und Constanze : mit obrigkeitlicher Bewilligung ; Einladungs-Billet ; hochzehrende Herren und Damen! Unterzeichnete Gesellschaft, hat die Ehre Sie zu benachrichtigen, daß sie mit ihren 3 1/2 Schu.
   Bretzner, Christoph Friedrich (Beitragende) 1818
   1 Plakat : sw ; 20 x 16 cm
   Druckschrift ; Theaterzettel
   Universitäts- und Landesbibliothek Düsseldorf

3. An sämtliche geehrtesten Herren Musik- und Theater-Freunde, Lieben und Musiker, zum neuen Jahre 1818
   Fischer, Joseph (VerfasserIn)
   1818
Specialized Information Service
Performing Arts

Sacchetto, Rita
Tänzerin, Sängerin, Schauspielerin
geboren 15. Januar 1880 in München
gestorben 18. Januar 1959 in Genoa-Nervi

Biografische Angaben
Gräfin Zamoyski; Geburtsjahr nach anderen Quellen: 1879 und 1888

Namensvarianten
Sacchetto, Margherita; Zamoyski, Rita

Weiterführende Links
- International Standard Name Identifier (ISNI)
- Filmportal
- Virtual International Authority File (VIAF)
- Wikidata
- Gemeinsame Normdatei (GND) im Katalog der Deutschen Nationalbibliothek
- Bayerisches Musik-Lexikon Online
- BNF Bibliothèque nationale de France
- DE-611 Kallope Verbundkatalog
- Wikipedia (Deutsch)
- Wikipedia (English)

Quelle:
http://d-nb.info/gnd/116713631/about (Lizenz: CC0 1.0)
via lobid-grid

... extended by fact sheets for agents and events
• The Specialized Information Service in numbers:

- ~800,000 Objects
  (Theatre bills, Photos, Videos, …)

- ~60,000 Persons
  (Actors, Dancers, Directors, …)

- ~6,000 Organizations
  (Ensembles, Institutions, Groups, …)

- ~60,000 Events
  (Festivals, Performances, Conferences, …)
The Challenges then and now

„The Laughing Audience and A Chorus of Singers“
Copperplate by William Hogarth, 1733
[Theatre Museum of the State Capital of Düsseldorf]
Typical challenges regarding the original metadata

- Different ways and frequency of delivery (mail, harvest, floppy disks, …)
- Different data formats and metadata standards
- Different scope and detail of description, no common vocabulary
- Little or no documentation
- Unstructured data / free text / “hidden information“
- Expectations vs. actual existing data
Raw data - challenges

Those challenges are basically the same as in 2016

• We face many of these challenges for each new data provider
• Many conversions and mappings are needed
  ⚠ potential loss of information
• Normalization, enriching and interlinking is needed
• Many small conversion steps that depend on each other
• Amount of data and steps to perform increases with each new data provider
• You can produce wonderful rich(er) data, but there is one thing to keep in mind: Giving back
How to give back?

Giving back to data providers

• Possibility to give back is very heterogeneous (various in-house systems, man power, financial situation, “mapping back”?)
• Take time to plan how to give back (which format/standard?) in close communication with the data provider
• Easy first step: hand data providers the results of your analysis
• Give out best practice recommendations (e.g. KIM)
• Make the data providers see the benefits
How to give back?

Giving back to the (tech or subject-specific) community

• Give out best practices
• Give out recommendations for tools
• Make code and documentation available
• Use mailing lists, ask questions, do pull requests
• Provide API / access
Workflow
→ „Behind the scenes“

„The Taming of the Shrew [IV]“
Set design draft by Traugott Müller, 1942
[Freie Universität Berlin, Institut für Theaterwissenschaft, Theaterhistorische Sammlungen]
Workflow in 2016

1) Analysis and normalization
2) Transformation to XML
3) Mapping to aggregation format EDM
4) Enrichment (entityFacts, geonames,…)
5) Deduplication (tbd)
6) Mapping to Solr-Indexformat

Advantage:
Step 4-6 is the same for all data
Workflow in 2019

What is still the same in 2019?

• Thorough **analysis** and **documentation** of delivered data is still the key step

• still following the principle of doing as many steps as possible for all **data** in the same way

• The wonderful world of XPath, XSLT and Xquery

• Europeana Data Model (EDM) as data model

• “Basic“ methods to normalize and interlink the data

• Still no deduplication, no API (yet)
Workflow in 2019

What has changed since 2016?

• Analysis step is partly automated now
• Mappings to EDM are “less clever“
  → clever steps are done later in the same way for all data
• Tools we use
  → especially to use of an XML-Database and a pipeline tool
• More modular
• Better performance :-)
Workflow in 2019

- currently ~200 tasks
- documents the workflow
- more modularity
- new providers are easily added
- easier to proceed from where it failed

XML-Database
- fast manipulations on each record
- great for analysis and visualization of huge collections
- supports JSON and CSV as well
Workflow in 2019

- favourite API for GND
- it is used in the fact sheets
- great for more complicated queries / facetting

- matching of “other“ authority data to GND via Reconciliation in OpenRefine with lobid-gnd
- results currently reviewed
Workflow

Raw Data

Analysis
- Understanding
- Feedback
- Documentation

Preprocessing
- Normalization
- Merging / Chunking
- Conversion to XML

XML

Mapping
- Map to EDM
- Parsing from free text to make the most of the given data

Enriching
- Enrich authority data via GND
- Match other entities to GND (half-automatic)

Indexing
- Index object data and authority data to Solr search engine

Other Sources

Authority Index

Title Index

Enriched EDM-XML

EDM-XML

not data provider-specific

data provider-specific
Still challenging

- There is still no common vocabulary that is used by our data providers but they are working on it with our help
- Uniquely identifying entities from literals automatically is prone to error
- Keeping up with updates and changes of tools, namespaces, …
- You can not make information magically appear when it is not there…

What would be nice to have?

- Natural language processing to extract more events and agents from the description fields
- Visualization
- API (a sparql endpoint would be nice)
Thank you!
Visit performing-arts.eu and give us your feedback!

Contact: Julia Beck | j.beck@ub.uni-frankfurt.de
Project leader: Franziska Voß | f.voss@ub.uni-frankfurt.de