FROM RAW DATA TO RICH(ER) DATA: LESSONS LEARNED WHILE AGGREGATING METADATA

Julia Beck, Frankfurt University Library

Introduction

The Specialised Information Service Performing Arts (SIS PA) is part of a funding programme by the German Research Foundation that enables libraries to develop tailor-made services for individual disciplines in order to provide researchers direct access to relevant materials and resources from their field. For the field of performing arts, the SIS PA is aggregating metadata about theater and dance resources from currently, mostly, German-speaking cultural heritage institutions in a VuFind-based search portal.

In this article, we focus on metadata quality and its impact on the aggregation workflow by describing the different, possibly data provider-specific, process stages of improving data quality in order to achieve a searchable, interlinked knowledge base. We also describe lessons learned and limitations of the process.

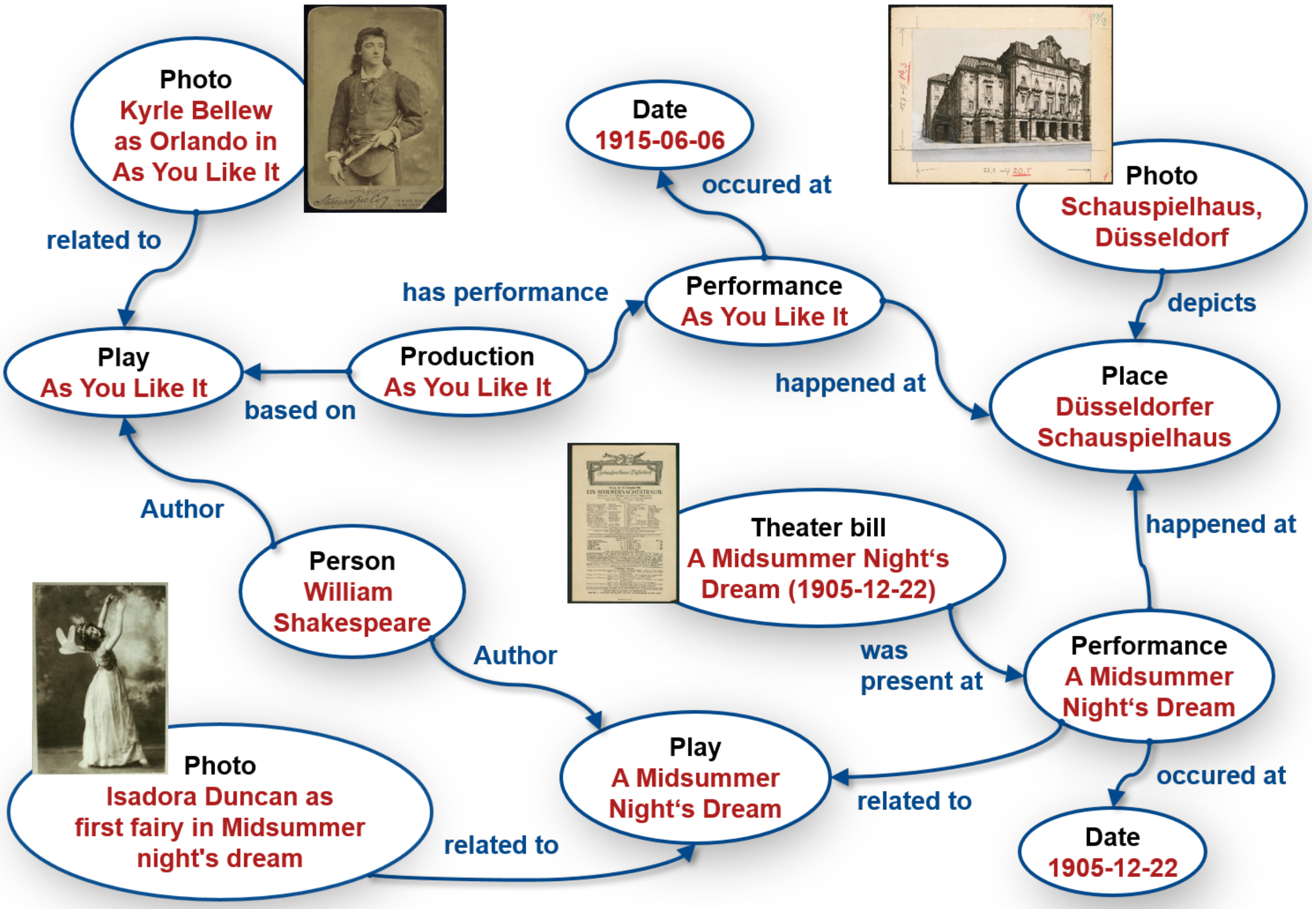

Figure 1. Schematic depiction of a knowledge graph in the performing arts domain. (CC 0)

Challenges in Aggregation

The metadata that is aggregated in the SIS PA origins in different branches, i.e. libraries, archives and museums, different types of performing arts such as theater, dance, puppetry or performance art and overlaps with other disciplines like filmmaking and music. Due to this variety, in which different standardizations and archiving workflows are preferred, the gathered metadata tends to be very heterogeneous in data format, metadata standard, data model and vocabulary.

Additionally, differences in scope and detail of description as well as the handling of entities yields further challenges for the aggregation process. Especially the description of performances as some data is described object-centric while other collections focus on the description of events. In order to give the user a preferably consistent search experience, it is important to normalize this heterogeneity and make the collections linked and searchable on common entities and terms.

Workflow

Starting from the originally delivered metadata, the current workflow to achieve aggregated linked data that researchers can access in the discovery system consists basically of four stages: (1) thorough analysis and documentation of the delivered data, (2) preprocessing of the original data, (3) modeling and transforming the preprocessed title and authority data to EDM for more interoperability, (4) postprocessing, interlinking and enrichment of the entities.

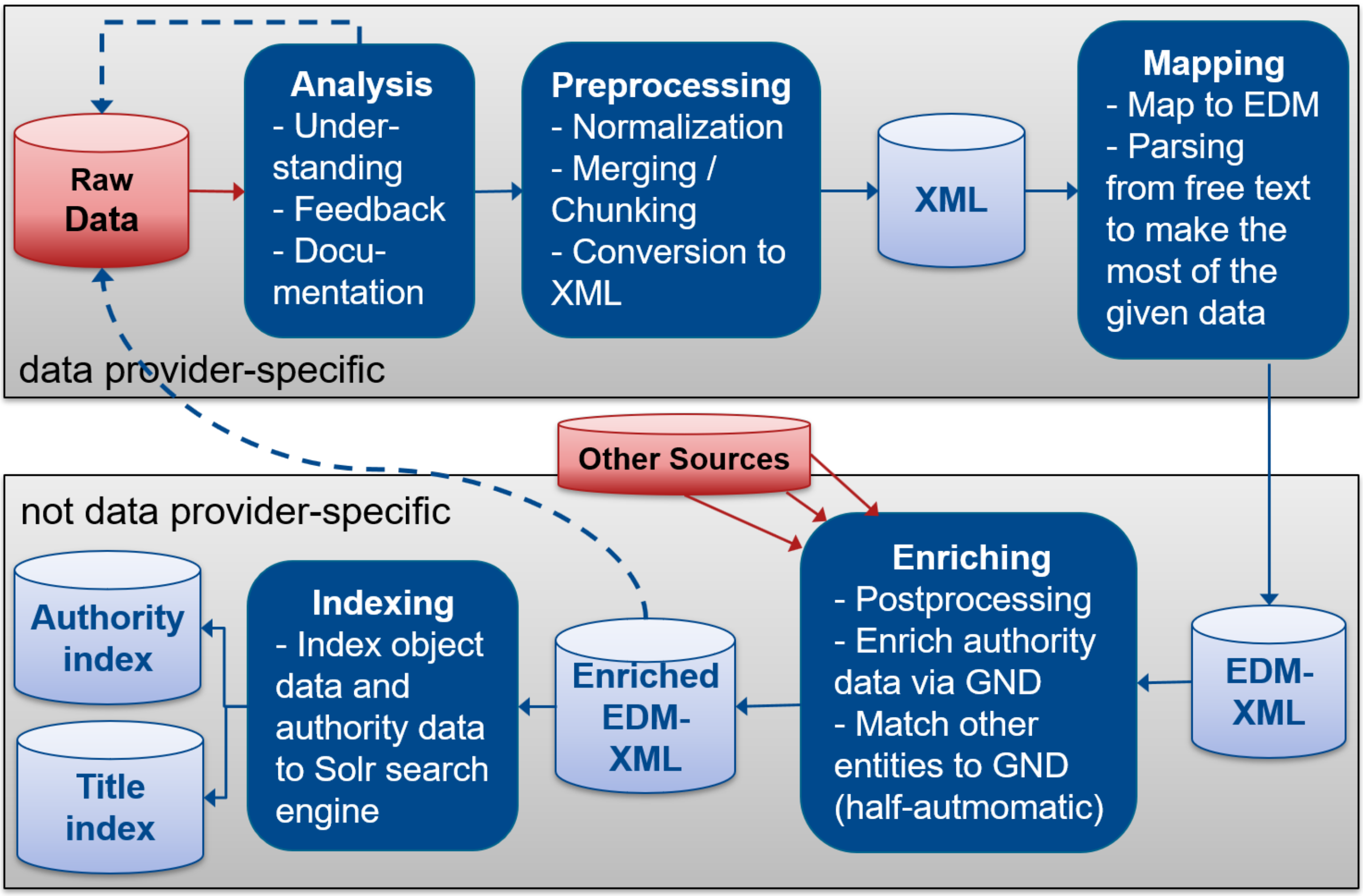

Figure 2. Data workflow from originally delivered data to enriched data that is indexed in the search engine. (CC 0)

1 – Data Analysis

The first stage involves thorough analysis of the delivered metadata in cooperation with the data providers to understand the data structure and given information. This step is key in improving the search experience for users because it determines where what kind of information is found in the collection. The metadata is validated to find possible problems like invalid data and duplicate or missing identifiers that need to be handled in-house. A short analysis report for the data provider is written that documents statistics and facts about the data. It further contains a first draft on how the data could be mapped into the SIS PA data model regarding the content of the data fields.

The analysis reports can also be used to find common structures and vocabularies in the data and generate best practice guidelines about performing arts related metadata. This knowledge is shared among others with the study group ARCHIV of the Society for Theatre Studies in Germany. In addition, we are following the activities in other projects on performing arts data like WikiProject Performing arts.

2 – Preprocessing

In the second stage, where most steps are data provider specific, the metadata is preprocessed which means that it is prepared to be mapped into the internal SIS PA data model which is based on the Europeana Data Model (EDM). Since most of the data is delivered in XML serialization and XSLT mappings from the Europeana community transforming XML files to EDM can be reused, all data that is not delivered in XML is converted to XML in this step.

Huge data files are split into smaller, more manageable batches while one-record-files are merged to larger batches. Depending on the original data quality and standardization, special characters like non-sort characters are normalized and certain free text data fields are parsed to make their content available for the search. Normalization is also needed if the content is not strictly following the rules of a certain metadata standard.

3 – Mapping

After normalization, the data is mapped into an extended version of EDM. Deeper insights into the extensions of the data model to make it meet the needs of performing arts can be found here. Besides reusing existing mappings from the community, several new mappings to EDM were written for individual data models following the findings of the analysis report. This transformation maps the original data fields into the corresponding EDM classes and properties on a very basic level. Further enrichment will be done in the next stage in order to be able to do as many enrichment and normalization steps as possible in the same way for all data providers.

Hierarchies and other relations between records are preserved or generated if possible. If a data provider delivers their own authority data that is not already linked to e.g. GND or VIAF, this data is mapped to EDM contextual classes like edm:Agent or edm:Event as well.

4 – Enrichment

In the fourth workflow stage, which is enriching, all data is gathered in an XML database and is normalized once more. Date values are mapped to UTC standard, while language values are mapped to ISO standard. The data is validated in terms of missing mandatory fields, collection hierarchy errors, plausibility checks for dates and other problems that might have occurred during the mapping. After the title records are normalized and validation errors corrected, authority records from data providers and authority identifiers used in the title data are gathered and stored in a separate authority database. This authority data is matched to the GND as it is widely applied in the German speaking community and some of the data providers make use of it already. By means of the lobid-gnd service that is integrated into Open Refine's reconciliation tool, authority records from data providers are matched to the GND and a project team member checks manually if the match was correctly made (s. Figure 3 as example for matching results). As authority records that do not exists in the GND yet are added during the project, the GND is gradually complemented with more agents and events from the performing arts domain.

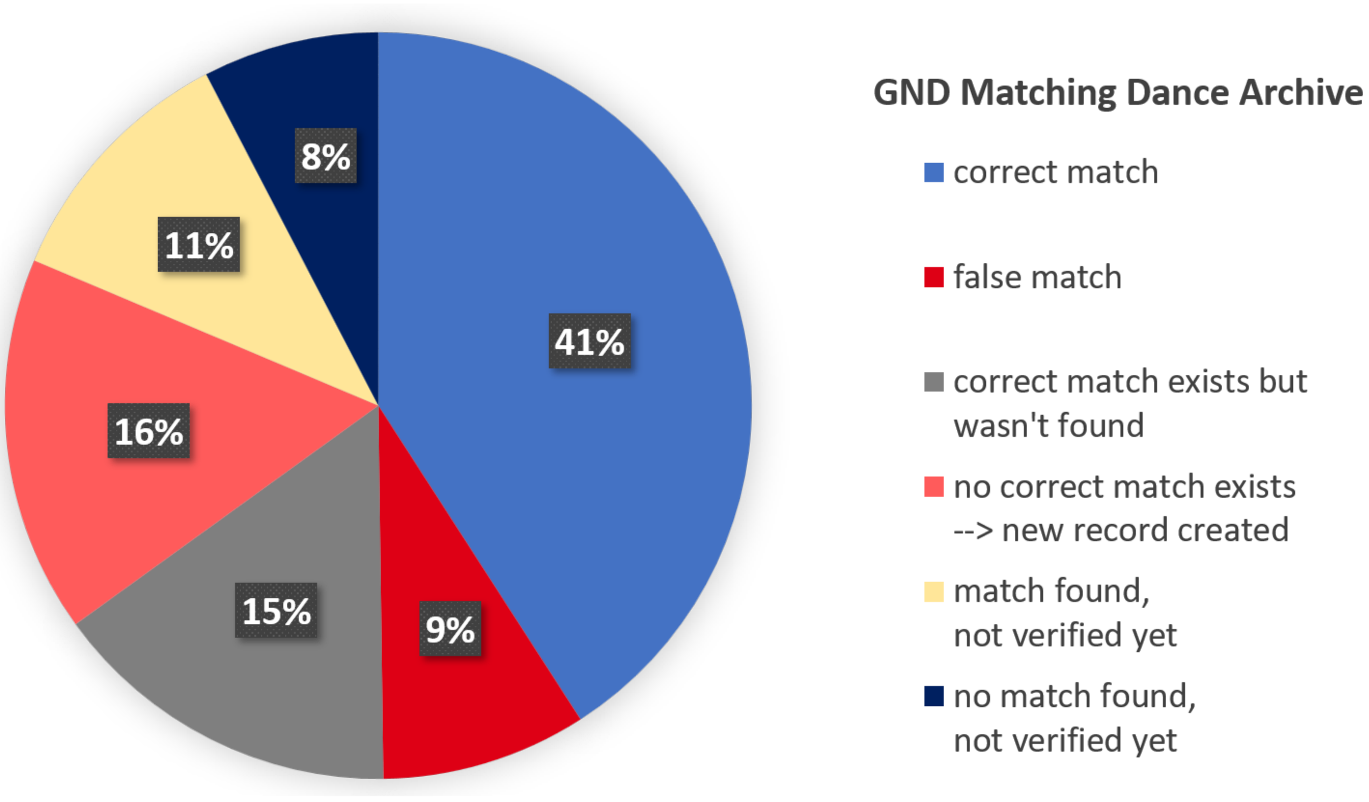

Figure 3. Results of the current GND matching task for the local agent database of the German Dance Archive. (CC 0)

For all GND identifiers used in the data, the GND entity dump from the German National Library is loaded to enrich the authority data. For example, alternative labels in the authority index help to improve search results and occupations that are based on a common vocabulary enhance faceting. In case the data contains GND identifiers that are no longer valid, these are replaced by the associated new GND and the GND type is corrected if it is not corresponding to the contextual class. After enrichment is done, the performing arts knowledge base can be used for various purposes. In the SIS PA, the title and authority data are indexed each into one search index core and are accessible through the project's discovery system.

In order to achieve metadata roundtripping, the results of the analysis and enrichment are given back to the data provider in form of simple .csv files. The files contain, if applicable, corrections and normalizations like the standardization of date and language values as well as concordance lists of local authority identifiers and GND identifiers. This process offers the possibility to enrich the original database of a data provider. Though, as many data providers do not have the capacity or technical support to automate the import of metadata, the modified data is often ingested manually if time allows which simultaneously allows for verification.

Lessons Learned

Thorough analysis of the delivered metadata in close cooperation with the data providers in order to make the data as useful and findable as possible is key in upgrading the search experience for users. By working together and exchanging results, the data quality can be improved both for the data provider and the aggregator.

Since the matching task from literals to entities is hard to automate retrospectively and prone to error, we noticed it is important to make data providers aware of this problem so that the in-house data preferably already contains authority data. We did not attempt to match literals without further information to authority data yet as the automatic matching approach on in-house created authority data already lead to errors that had to be corrected manually.

We also learned that the use of database and pipeline tools can improve the workflow substantially. XML databases like BaseX that are made for the task of managing huge amounts of XML can reduce the process time immensely while pipeline tools like Luigi can help to keep track of all the small steps of the process.