Noname manuscript No.
(will be inserted by the editor)

Performing Entity Facts

The Specialised Information Service Performing Arts

Julia Beck · Michael Büchner · Stephan Bartholmei · Marko Knepper

Received: date / Accepted: date

Zusammenfassung In a theatre play, persons appear as playwright, director, actors, etc. The play may have several performances with changing casts while actors may contribute to other plays and the role of a contributor may vary. Persons and the character of their contribution are a major focus in the performing arts domain. In order to create a domain specific and comprehensive research portal, current indexing techniques are combined with linked data methods giving access to person related information. The *Specialised Information Service Performing Arts* aggregates numerous metadata sources documenting the holdings of German-speaking cultural heritage institutions. This information source is extended by links to the recently established service *Entity Facts* by the German National Library that provides details to the persons related to the metadata records. The portal is based on the VuFind framework with index files that cover the metadata of all data providers and the cached data of all related authority records from *Entity Facts*. In order to achieve this, the standard model of VuFind MARC21 has been replaced by the Europeana model EDM. This allows for modeling all data – metadata and authority data – following linked data principles. While the respective mappings could be re-used new record drivers and indexing rules had to be defined.

Schlüsselwörter Linked Open Data · Europeana Data Model · Performing Arts · VuFind

J. Beck
University Library Frankfurt am Main
Tel.: +49-69-798-39387
E-Mail: j.beck@ub.uni-frankfurt.de

M. Büchner
German National Library, Frankfurt am Main
Tel.: +49-69-1525-1774
E-Mail: M.Buechner@dnb.de

S. Bartholmei
German National Library, Frankfurt am Main
Tel.: +49-69-1525-1783
E-Mail: S.Bartholmei@dnb.de

M.Knepper
University Library Mainz
Tel.: +49-6131-39-32895
E-Mail: M.Knepper@ub.uni-mainz.de

1 Introduction

At the University Library Frankfurt am Main, the *Specialised Information Service Performing Arts* is developing the search portal http://performing-arts.eu for academic research on theatre and dance studies. The target audience are scholars who conduct research by means of primary sources on performing arts. In close collaboration with experts in the discipline, the overarching catalogue is aggregated with metadata about the most significant performing arts-related collections of libraries, archives and museums in the German-speaking world.

When the web portal launched its beta version in June 2016, about 330.000 records from ten participating cultural heritage institutions[1] were searchable. Users can find metadata about performance-related objects like prompt books, costume blueprints and other material that is part of the creation of and the performance itself as well as information about the persons involved.

[1] A list of participants and cooperation partners can be found on http://performing-arts.eu/spages/netzwerke. Accessed 10 October 2016

2 Aggregating the Metadata Pool

The gathered metadata turned out to be very heterogeneous and domain specific due to the range of different cultural heritage institutions involved and the resulting variety of different data acquisition workflows. This could not be compensated by requiring metadata standardization at the time of delivery as the data providers' technical infrastructure was very divergent. Therefore, individual technical support and flexibility in the receipt of metadata was necessary.

The delivered metadata included rich and internationally common metadata standards like LIDO from the museum sector, MARC21 and the proprietary format PICA from the library sector as well as METS/ MODS which is used for digitalized material in libraries. Furthermore, metadata from the archival Allegro and FAUST database systems and from individually configured or implemented systems was received as XML-, SQL- or CSV-exports.

Identifiers from the authority data of the German National Library (GND) for persons and organizations make it possible to identify a person or organization via a globally unique identifier and to get more information about them. This basic concept of Linked Open Data to make use of unique identifiers, enables cultural heritage institutions to benefit from the re-usable and extendable metadata that is gathered around the world [1].

But only a few metadata sets like the digital collection *Düsseldorfer Theaterzettel*[2] and the *Komplex Mauerbach*[3] collection from the Don Juan Archiv Wien did comprise GND identifiers, while other data sets had no authority data. Though, some of the data providers offered additional biographical data about persons like the date and place of birth within their data model. By matching the name and the dates of birth and death to information in the Swiss Theatre Dictionnary which had been enriched with GND identifiers by a librarian in the course of the project, the Swiss Theatre Collection was enriched with GND ids.

3 Data Model and Architecture

In order to realize a state-of-the-art search interface, the discovery system VuFind [12] was used as the central framework. The search interface runs on the open source search engine Apache Solr that offers great performance and scalability and includes features like fault tolerance, ranking and facets. All other functions of the portal, such as the news stream and the static area, were integrated into the VuFind installation and implemented as modules. This approach allows for the seamless integration and the re-use of the additional code which will be made available to the community.

For one of the project's aims was to make the metadata available for reuse as Linked Open Data, the heterogeneous metadata had to be normalized and transformed into one aggregation model. The library metadata standard MARC21, which is supported by VuFind would not have been appropriate to map museum and archival metadata and to model Linked Open Data. In addition, the possibilities to adapt and extend the data model to meet the domain's requirements in representing theatre or dance premieres with date, place and involved persons are very limited.

Hence, a universal and flexible metadata standard, the Europeana Data Model (EDM), was selected as data model because it meets the requirements of all different types of heritage institutions and may be directly exported as Linked Open Data [8,10,13]. The inheritage principle of the properties makes this standard flexible without giving up the interoperability. In addition, the advantage by re-using existing mappings [6] exceeds clearly the effort for the VuFind integration. The entire aggregation workflow from the original data models to their display in VuFind is modelled in Figure 1.

In order to cover all necessary properties in the performing arts domain, extensions of the DM2E[4] project and specific properties from the ECLAP[5] vocabulary have been added to EDM as depicted in the example in Figure 2. Relevant extensions include the different roles of the contributors to a (handwritten) document such as pro:author, bibo:recipient and pro:translator from DM2E. In order to reflect the specific demands in the performing arts domain, roles of contributors to a theatre play like eclap:actor, eclap:dancer and eclap:director from ECLAP have been added. Further domain specific properties include eclap:perfomancePlace and eclap: performingArtType.

While the DM2E ontolgy has been defined as an extension of the EDM model making use of the inheritance principle for properties, the ECLAP model has an independent structure reflecting the specific needs and the structure of the data of the ECLAP project, especially the social graphs [4]. In the project described here, it was possible to model the collected data by adding properties preserving the inheritance principle. This allows for consuming the data by third party systems in different semantic grains. In addition, the

[2] http://digital.ub.uni-duesseldorf.de/theaterzettel?lang=en

[3] http://www.donjuanarchiv.at/historischer-bestand/komplex-mauerbach.html

[4] http://dm2e.eu/

[5] http://www.eclap.eu

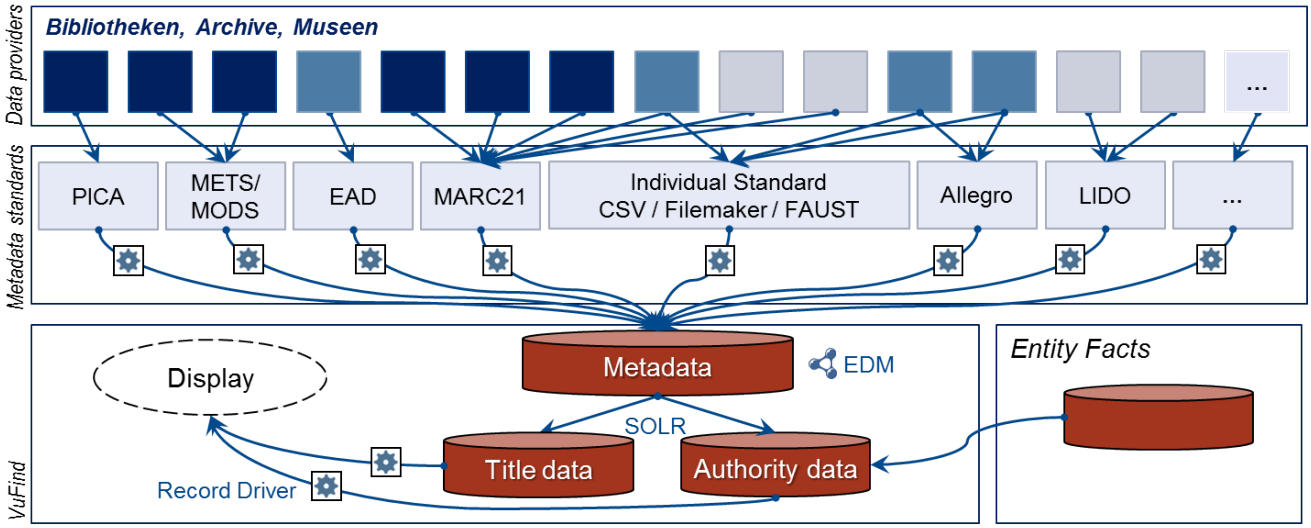

Abb. 1 Aggregation model: The provided metadata is mapped from its original data models to EDM. The resulting metadata is indexed into SOLR in a title data and an authority data core respectively. The authority data is enriched by means of the *Entity Facts* Service. For displaying the metadata in VuFind, a EDM record driver is needed.

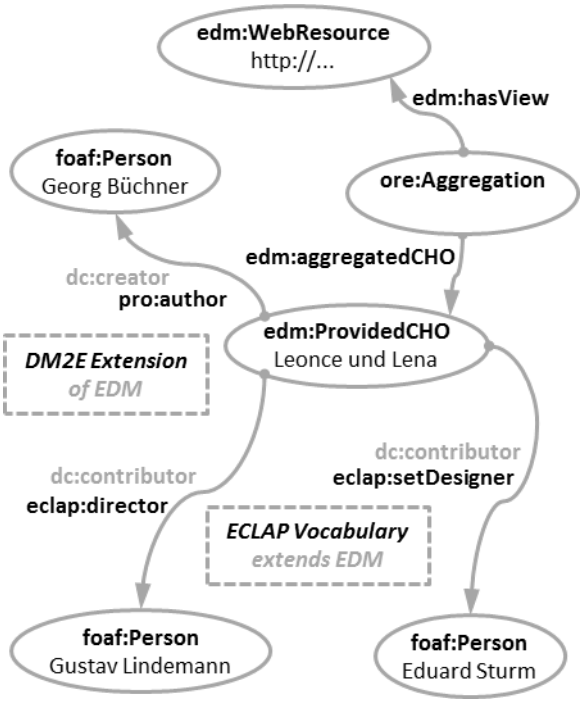

Abb. 2 Example demonstrating the use of the data model. The DM2E extension and the additional ECLAP vocabulary provide more specific roles for persons as needed in the perfoming arts domain.

re-use of EDM mappings was possible with minor modifications. However, this is a pragmatic approach and if the complexity of the data increases, it is likely that the necessity to develop the EDM base model will arise.

In contrast to the more Linked Data oriented approach in [9] by using a triple store, in this project the XML serialization of the EDM data was used to match seamless into the record oriented concept of the VuFind framework. The metadata model of VuFind was exchanged by defining new mapping rules for the incorporated SOLR index in an XSLT sheet. During the index creation, it was ensured that subject headings remain original and domain specific to help users to find the searched information. Though the subject headings will have to be mapped to a Linked Open Data ontology to merge spelling differences and semantically equivalent content. Furthermore, a new record driver that is responsible for the proper display of the metadata of a record in VuFind was implemented in PHP.

During the project, the importance of the people and organizations involved in the creation of and the performance itself became obvious. Information about agents was gathered in VuFind's SOLR search engine in form of an additional index core [11]. It contains edm:Agents that are constructed by mapping the original metadata to appropriate properties in edm:Agent. Agents are either identified by the GND identifier or by a generated identifier in case it is an agent that could not be matched to a GND identifier. By means of the GND identifier, it was possible to enrich this metadata with information provided by the *Entity Facts* service.

4 The Entity Facts Service

Entity Facts[6] is a data service run by the German National Library. It provides ready-to-use "fact sheets" on

[6] http://www.dnb.de/EN/entityfacts

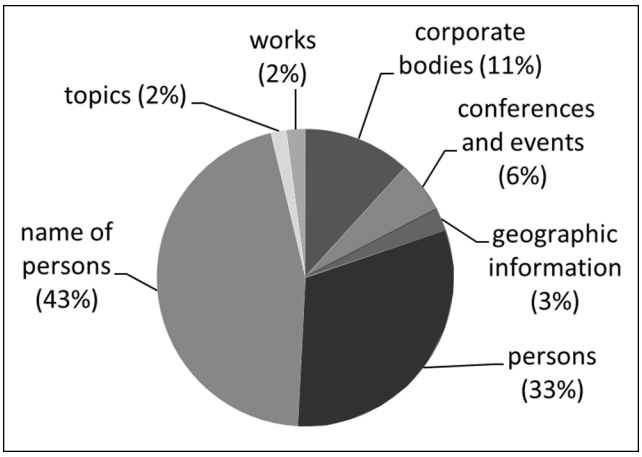

Abb. 3 Type of entities within the GND authority file (13.5m entities in total in August 2016)

the entity types "person" and "corporate body" of the Integrated Authority File (GND).

The GND itself is an authority file for persons, works, corporate bodies, conferences and events, geographic information and topics (Figure 3) which was initially used for the cataloguing of literature in libraries of the German-speaking countries. The authority file is hosted by the German National Library and run cooperatively by a library network, the German Union Catalog of Serials (ZDB), the Swiss Library and numerous other institutions. In the last couple of years the GND has become more and more popular, so it is now increasingly being used by institutions of other cultural sectors like archives, museums and monument protection agencies as well as by different projects or web applications.

Due to the amount of data, it was not easy to access the data of GND in the past. The large and extensive RDF dumps are not lightweight, they are expensive to parse and difficult to process. The GND also contains a lot of less relevant information which is not useful for most applications. Another problem is that the RDF data of GND is not suitable for presentation. Data records have a specific format (e.g. ISO-formatted dates) which cannot be used for presentation unless they are processed beforehand.

Developing and publishing *Entity Facts* was one step on the way to increase the visibility and possibility of reuse of GND's CC0 1.0-licensed data. The idea of *Entity Facts* is to provide a very easy, lightweight and intuitive to use data service which has ready-to-display JSON-LD data (e.g. all dates are formatted for a human consumption: "August 28, 1749"). *Entity Facts* is always up-to-date, all edits and additions made for the GND will appear on the fly in the data service as well. Table 1 lists exemplary information about Johann Wolfgang von Goethe provided by *Entity Facts*.

Another goal was to enrich the data of the GND by adding external information. A good example is the enrichment with a person's photo or a logo of a corporate body in *Entity Facts*' data field "depiction" (URI: http://xmlns.com/foaf/0.1/depiction). After analyzing various databases, it was concluded that Wikidata[7] provides suitable information to enrich GND's data in the context of the data service.

Wikidata is a free knowledge database with information collected and collated by a community of volunteers. The intention of this Wikimedia Foundation project is to provide a common source of structured data which can be used by anyone under a public domain license. Wikidata is providing a data dump every week, which contains various entities also described within the GND. Many items of Wikidata link to the GND (property P227) and also have a picture (property P18) as well as much other information which are suitable for an enrichment of the GND.

Interlinking to other resources which deal with facts about an entity was another goal of *Entity Facts*. The idea was that external applications could use *Entity Facts* to link to other websites and portals with more detailed information or which have another point of view regarding a specific entity. Examples for this "sameAs" relation (URI: http://schema.org/sameAs) are the Virtual International Authority File (VIAF), the Internet Movie Database (IMDb), the authority file of the Bibliothèque nationale de France (BNF) or of the Library of Congress (LC authorities).

The enrichment of the interlinking was implemented by using BEACON files [2]. BEACON is a simple file format with links to webpages, portals, data services etc. which have content relevant to entities of authority files. Those BEACON files are provided by external organizations[8].

The registered BEACON files with information on persons provided by third parties are harvested regularly. The contained links are added to the *Entity Facts* authority data. All following requests to the *Entity Facts* service include the owl:sameAs statements enriched with links to the third party resources.

The first version of the *Entity Facts* service was published in March 2014. It included and provided information from the GND on the entity "person" and linked to other data sources [5]. Those links are based on publicly provided BEACON files evaluated by the German National Library. Since then, *Entity Facts* has been continuously improved: The data service now al-

[7] https://www.wikidata.org

[8] A small database of available BEACON files for mostly German content can be found under https://de.wikipedia.org/wiki/Wikipedia:BEACON

Tabelle 1 Examples of information within an *Entity Facts* data sheet (Johann Wolfgang von Goethe, http://hub.culturegraph.org/entityfacts/118540238)

Data field URI	Example for a person	Description
http://d-nb.info/standards/elementset/gnd#preferredName	Johann Wolfgang von Goethe	Complete preferred name of an entity in a specified format
http://d-nb.info/standards/elementset/gnd#variantName	Johann Wolfgang Goethe, J. W. von Göthe, ...	Array of variant names in a specified format which are also used for the entity
http://d-nb.info/standards/elementset/gnd #professionOrOccupation	Schriftsteller, Publizist, Politiker, ...	Array of professions and occupations connected with the entity/person
http://d-nb.info/standards/elementset/gnd#placeOfBirth	http://d-nb.info/gnd/4018118-2 (Frankfurt am Main)	Place of birth as GND URI and human-readable label
http://d-nb.info/standards/elementset/gnd#dateOfDeath	22. März 1832	Date of death as human-readable label
http://d-nb.info/standards/elementset/gnd#relatedPerson	http://d-nb.info/gnd/118607621 (Friedrich Schiller)	Array of related persons of the entity as GND URI and human-readable label

so provides "fact sheets" for the entity type "corporate bodies" and enriches all entities with pictures (photo, painting etc. of a person or the logo of an institution) if they are available on Wikidata. *Entity Facts* runs on a flexible and stable infrastructure, which makes it easy to extend the data service and the database, too.

The *Entity Facts* service was co-developed by the German Digital Library (DDB), Germany's national cultural heritage data platform and portal (https://www.ddb.de). DDB's entity pages[9] serve as the reference implementation for applications using the service.

5 Research Interface

Making use of the *Entity Facts* service results in the enrichment of the archival material in the *Specialised Information Service Performing Arts*. Contributing and other related persons are linked to fact sheets with more information about these persons' life and work in case a corresponding GND identier is available. Furthermore, depictions of persons and organizations can be displayed. Particularly interesting regarding the performing arts domain, is the linking to the Internet Movie Database. Also, the generated index based on name entries allows for the search in the set of all persons that are relevant and have been identified. The enrichment via *Entity Facts* will improve this search for persons as it adds spelling variations of names to the index.

Generally speaking, the generation of the search index and the display of all records is based on the aggregated metadata after normalization and transformation. As a consequence, any information loss in the process of normalization and transformation had to be minimized. Especially dates and places needed normalization as standardization was often missing. Due to the lack of a commonly used vocabulary in the domain, the normalization of subject headings as mentioned above, the normalization of material types like playbills and kinds of photographies as well as the kind of contribution a person does to a performance turned out to be a challenge that required thorough data analysis and expertise by the data providers and professionals in the domain. This was also because of the different depth of description that made it hard to map and combine information within the same facet without information loss.

The resulting advantage of the normalization is the homogeneous display of the records for each object and a straightforward generation of the index that is used for the facets. The research interface presented to the user allows for the search of material about certain plays and performances which the user can filter by data provider, material type, collection or date. The user can choose a record from the result list and gets a more detailed view with the title, a description of the object, its date and place of creation, involved persons and more advanced information like links to related performances or a link to the associated collection or series if this information is given. For example, if a user searches for photographies of a performance of Georg Büchner's "Leonce und Lena", the user can type the name of the play, filter the results by material type and retrieves suitable photographies with information about the depicted persons and, if available, links to the related performance and other objects that belong to the same performance.

A similar approach to make metadata about performing arts searchable, is the ECLAP portal which was funded by the European Commission until 2013. The ECLAP namespace, that the described system de-

[9] e.g. https://www.deutsche-digitale-bibliothek.de/entity/118540238, the person entity page of Johann Wolfgang von Goethe

ploys, originates from the mentioned ECLAP project and counteracts the lack of a comprehensive vocabulary in the performing arts domain [3]. ECLAP's social graph offers a comprehensible view on the data and focuses on information about contributors and their relation to a play, a performance or another contributor. By contrast, the *Specialised Information Service* focuses on additional information of a person's life, date of birth or links to other information resources such as wikipedia.

Another example for a search portal in the performing arts domain is the *AusStage*[10] project which has a focus on live performances in Australia and collaboratively collects information from users like artists, researchers, librarians etc. Furthermore, the Swiss Theatre Collection and the Swiss Dance Archive are planning a joint platform [7] for performing arts based on Linked Open Data.

6 Conclusion and future work

The *Specialised Information Service Performing Arts*, funded by the German Research Foundation, is a successful example how new research resources are made available by combining the advantages of the semantic web with a state-of the-art search interface on the base of real data.

The EDM model has proven to be a flexible and expendable basis. More and further enriched data from the data providers and the ongoing progress of the project will show how the data model needs to be developed. The successful integration of the *Entity Facts* service still has potential by using additional person properties for the retrieval interface. However, the practical use will depend on the efficiency of enrichment with contextualisation identifiers on the data sets that have originally been delivered without authority data.

Future developments of the *Entity Facts* service will see the integration of *Entity Facts* into the Linked Data Service of the German National Library, adding the remaining GND entity types (e.g. "geographic information"). We will also improve and expand the JSON-LD data model and – of course – include even more data sources as links. Entity Facts will also enlarge the scope of the enrichments and for each enrichment it will clearly state its source.

Literatur

1. Baker T, Bermès E, Coyle K, Dunsire G, Isaac A, Murray P, Panzer M, Schneider J, Singer R, Summers E, Waites W, Young J, Zeng M (2011) Library Linked Data Incubator Group Final Report.https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/. Accessed 10 October 2016
2. BEACON link dump format, Format specification; July 6, 2014; http://gbv.github.io/beaconspec/beacon.html. Accessed 14 October 2016
3. Bellini P, Nesi P (2013) A Linked Open Data Service for Performing Arts. Second International Conference, ECLAP 2013, Porto, Portugal, April 8-10, 2013, Revised Selected Papers, pp 13-25
4. Bellini P, Nesi P (2015) Modeling performing arts metadata and relationships in content service for institutions. Multimedia Systems 21: 427-449. doi: 10.1007/s00530-014-0366-0
5. Böhme C, Büchner M (2014) Entity Facts - A light weight authority data service. SWIB14 – Semantic Web in Libraries. http://swib.org/swib14/slides/buechner_swib14_11.pdf. Accessed 14 October 2016
6. Charles V (ed) Olensky M (ed) (2014) EDM mappings refinements and extensions. http://pro.europeana.eu/taskforce/edm-mappings-refinements-and-extensions. Accessed 10 October 2016
7. Estermann B (2016) Big Data: Den digitalen Wandel aktiv gestalten. arbido 3/2016, pp 26-31
8. Europeana Data Model Documentation. (2016) http://pro.europeana.eu/page/edm-documentation Accessed 10 October 2016
9. Hatop G (2013) Integrating Linked Data into Discovery. Code4Lib J, Issue 21. http://journal.code4lib.org/articles/8526. Accessed 10 October 2016
10. Heath T, Bizer C (2011) Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, pp 1-136. 1st edition, Morgan & Claypool, San Rafael, Calif. doi: 10.2200/S00334ED1V01Y201102WBE001
11. Katz D, LeVan R, Ziso Y (2011) Using Authority Data in VuFind. Code4Lib J, Issue 14. http://journal.code4lib.org/articles/5354. Accessed 12 October 2016
12. Katz D, Nagy A (2013) Solr Power in the library. Library Automation and OPAC 2.0: Information Access and Services in the 2.0 Landscape, pp 73-99. doi: 10.4018/978-1-4666-1912-8.ch004
13. Mitchell E (2014) Library Linked Data. 5th ed. American Library Association, Chicago

[10] http://ausstage.edu.au