JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

JLIS.it

Towards an open and collaborative Authority Control

Barbara Katharina Fischer[a] with the cooperation of Jürgen Kett[b], Sarah Hartmann[c], Mathias Manecke[d]

a) Deutsche Nationalbibliothek (The German National Library)
b) Deutsche Nationalbibliothek (The German National Library)
c) Deutsche Nationalbibliothek (The German National Library)
d) Deutsche Nationalbibliothek (The German National Library)

Contact: Barbara Katharina Fischer, b.k.fischer@dnb.de
Received: 25 June 2021; **Accepted:** 23 July 2021; **First Published:** 15 January 2022

ABSTRACT

As digital transformation is speeding up, the need for a reliable retrieval is too. Libraries have long used *authority files* to enhance the search for information. Now, as the entire GLAM field is increasingly presenting its content online, national libraries face the requirement to provide authority data as reference points to a far more diverse community. The request is not limited to persistent identifiers but new records on non-librarian entities are needed. The German National Library (DNB) aims to provide an open framework that allows *collaboration* on all levels: editing the records, defining the regulations and standards plus ease the data flow in both directions. To this end, the DNB has started an ambitious project transferring both the authority file records and their regulations into a *Wikibase* instance. The article relates the findings working with the beta version of the software that drives *Wikidata*. To spur the process the DNB co-published the WikiLibrary Manifesto together with Wikimedia Deutschland. The institutions signing the manifesto shall cooperate to improve the building of a technical infrastructure that will ease knowledge equity through the *FAIR Data Principles* and the creation of a structured data ecosystem. The manifesto was signed by IFLA in June 2021.

KEYWORDS

Library; Wikibase; Authority control; Fair data; Semantic web.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

"What really distinguishes us is the way in which we collaborate on a major scale."[1]

When people discuss topics with verve and persistence, this is generally a sign of dedication and connection. The topic of "Opening the GND" features these positive qualities. It affects and moves many people. It raises questions on the major topic of collaboration, both in great detail and a vast range of different contexts. The opening quote to this article is taken from the historian Yuval Noah Harari's[2] much-acclaimed graphic novel "Sapiens", which tells the history of how humankind developed. It also describes our work in the *Office for Library Standards* (AfS) at the German National Library. Organising collaborations is at the heart of what we do. Our task is to facilitate the cataloguing of knowledge resources across national and disciplinary boundaries. We organise collaborations by promoting consensus on standards that we ultimately use to describe the world while keeping them equally comprehensible for all. Using these standards, the community of German-language libraries is defining how publications should be described with greater precision than by means of natural language so that others can definitively refer to them. This is where the *Integrated Authority File* (GND) comes into play. Harari refers to the nature of humankind as a whole, and how this differs from the character of chimpanzees, for example. The work of cataloguing, the definition of media based on the rules of descriptive and content cataloguing, is far removed from the challenges faced by Homo Sapiens during the Stone Age. And yet, in a sense, it is simply a different section of the same light beam. As a result of the "cognitive revolution"[3] that occurred back then, today, we are facing the challenges of the digital transformation. And this too we will master precisely because of our ability to collaborate. This is what we do.
In the course of opening the GND to include communities beyond library institutions, one thing has become ever clearer: the GND is much more than just a collection of nine million authority data records on people, places, corporations, conferences, works and subject headings.[4] It also describes an organisational structure that reflects the state of its current users. It refers back to a certain data model that is based around the needs of its users. It is subject to specific rules and can be regarded as a specialist tool within an specialised environment defined by the requirements of the library community. Yet the new user groups are organised differently. They have other data models. They catalogue the objects of their interest according to different rules and use a different technical infrastructure. And yet they are still very interested in using the GND authority data. They don't just wish to use the identifiers in their cataloguing work, but also want to be able to create new GND data records when they see a need to do so. They want to become an active part of the GND community. To this end, we need to work together to consider carefully what we can change, and how much, without damaging the core of the GND. This is because everyone wants to preserve its reliable quality. Our task is once more to organise our collaborative efforts in line with a collective intentionality.

[1] Quote taken from Harari 2020, p. 68.
[2] Harari 2020.
[3] On the concept of the "cognitive revolution", cf. Harari 2012, pp. 11-100.
[4] The record type *Conferences* in the GND makes particularly apparent how interwoven the GND is with its users in the world of libraries. That is because this record type describes a specific kind of publisher. See all categories in the GND ontology: https://d-nb.info/standards/elementset/gnd

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

An instrument for broadening participation

Opening the GND is like the concert given by an entire orchestra of stakeholders and activities. One instrument in this orchestra, a starting point for a careful adaptation, is the technical environment in which the GND is rooted. It is not the notion of dispensing with the existing technical infrastructure, but much more the idea of offering a parallel infrastructure, that has drawn our attention to the database software Wikibase[5]. Wikibase is a piece of open-source software from the Wikimedia Foundation. This foundation has previously developed the Mediawiki software, which is used to operate millions of Wikis around the world. The most famous Wiki is Wikipedia, operated by Wikimedia. The Wikidata project was launched nine years ago with the aim of improving Wikipedia. A database for structured data with which one can describe the world in a way that can be read by both humans and machines alike. The software empowering Wikidata is Wikibase. Wikibase features certain properties designed to make large-scale collaboration easier:

- It offers web-based access.
- It facilitates parallel collaborative working.
- It automatically logs the version history and its editors.
- It offers a dedicated discussion page for every data record.
- It is geared towards multilingual user communities.
- It offers a simple and flexible (though also limited) data model.
- Entering new content works easily and intuitively.

We intensively studied these properties at the German National Library in 2019 and summarised our conclusions in our evaluation[6] in collaboration with Wikimedia Deutschland. In this context, we also explored current weaknesses in the system and potential areas for development. In its current iteration, the system falls far short of meeting all the requirements for an ideal editing system and hub for cultural institutions. To this end, it still needs to outgrow its origins as a piece of Wikidata software. Nevertheless, we were able to identify the fundamental prerequisites for its productive use in the context of the AfS. What matters is less the current status of the product and more the inherent potential in its further development and the establishment of a broad community in the cultural sector.

In 2020, we first considered how to make the most effective use of Wikibase in broadening participation in the GND, before creating the conditions for implementing our plans as efficiently as possible. We decided to become active on three levels. We want to:

- Create a second home for the GND as an authority file within a Wikibase database. New user communities can make suggestions for new GND data records more easily and independently of the existing technical structures, and compare their data to the GND with greater ease in order to avoid duplication.
- Create partnerships with Wikimedia and other institutions also wanting to use Wikibase, in order to collaborate on improving the software so as to ultimately establish an ecosystem for cultural data and research data.
- Thirdly, we want to re-order the very frameworks underpinning the GND and our cataloguing work, make these more accessible and easier to adapt to any changes.

[5] Link to the Wikibase website: https://wikiba.se/

[6] Link to the blog post on our evaluation: https://wiki.dnb.de/pages/viewpage.action?pageId=167019461

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

The second home of the GND

In the world of libraries, the GND has long served as a referencing and rationalisation tool, as did the four authority files that preceded it. It is integrated into certain frameworks and proprietary software structures that are, however, relatively inaccessible to users from outside the world of libraries. We believe that we can use Wikibase to make it easier for some of these target groups to collaborate on the GND.

To this end, we wish to import all the existing GND data records and their corresponding links into a Wikibase entity in 2021. This may sound like a simple task. However, Wikibase's importation interfaces are still very much aligned with the needs of Wikidata. For this reason, we have sought professional support from a Wikibase specialist, who is assisting us as a service provider in the transfer of the database infrastructure, the data importation and the creation of user-friendly input screens. In the next step, we will then invite experienced and new GND users to test the data-entry and search processes in the new environment so that we can further improve these.

During the second half of 2021, we are planning a technical workflow for synchronising the GND Wikibase entity with the CBS system[7]. The plan is to enable new and existing users without any WinIBW[8] access to enter their data as a suggestion in the Wikibase entity.

One long-term goal is to offer a user-friendly and supportive data-recording environment for the GND. The *GND web forms*[9] represent a first step in this direction, as they are considerably more user-friendly than the data-entry systems used by libraries. The web forms currently can be used to record people and corporations. However, this currently envisaged approach is not flexible enough. In addition to the two aforementioned GND record types, there are four more. These six record types unite approximately 50 entity codes[10], each with specific properties via which the respective entities can be recorded as GND data records. These would require a dynamic entry form that adapts to the entity type or usage context chosen, offers necessary and typical entry elements, highlights useful entries and thus guides the user through the entry process. It remains to be seen whether Wikibase represents the right platform for this in the medium term. At present, Wikibase lacks such features. For now, no update to the generic entry interface is planned "ex works". There is also no option of limiting the offering to fundamental elements or values. The user is always confronted with the full range of properties and values, and isn't offered any assistance in decision-making. One aim for 2021 is to establish whether this can be facilitated via the development of a Wikibase expansion, and also which changes would have to be implemented in Wikibase by Wikimedia in order to more adequately support the creation of customisable entry forms that assist the user.

[7] CBS: proprietary library data-entry software from OCLC.

[8] WinIBW: licensed software for entering data in the GND.

[9] The GND web form for persons and corporate bodies is specifically intended for users from cultural institutions such as smaller libraries, archives and museums who would like to create or modify small quantities of data records in the GND. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_Webformular/gnd_webformular.html

[10] Details on entity coding in the GND https://wiki.dnb.de/download/attachments/90411323/entitaetenCodes.pdf

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

The WikiLibrary Manifesto

Another area our work will focus on in 2021 is our partnership with Wikimedia Deutschland and other institutions in order to improve Wikibase as a technical infrastructure. Adherence to the FAIR Data Principles (Findability, Accessibility, Interoperability and Reusability)[11] when providing data is becoming increasingly important in an ever-growing number of contexts. Data are to become more interlinked in order to make it easier overall to generate new knowledge. This especially applies to data that were generated using public funding. This represents a great challenge for many institutions. It raises the question as to whether they should offer their data in a collective pool for structured data, like data portals. Such institutions must ask themselves whether they are willing to face all the potential consequences, such as sacrificing control over the data model, data-recording rules and quality-assurance processes. Or should they instead use stand-alone solutions and thus accept that their data will be less visible and get re-used less? By broadening participation in the GND, we wish to create alternatives. We are committed to creating a reliable, machine-readable and communally managed Linked Open Data Network for the arts, sciences and culture as a viable basis for FAIR knowledge. Instead of a central platform, we favour an open network of interlinked databases. This requires a communal organisational framework. We wish to provide this within a single network. A network is only ever as good as the partners within it. To this end, the German National Library co-published the WikiLibrary Manifesto together with Wikimedia Germany. Allmost forty institutions have already accepted our invitation. The manifesto invites the undersigning institutions to collaborate on the basis of the following principles:

- Promoting free licenses for data and their software environment.
- Shaping spaces where diverse communities thrive. (Community gardening).
- Providing structured data based on FAIR data principles in order to be able to transparently transform data into information to create FAIR knowledge.
- Promoting common core standards created consensually and collaboratively.
- Providing open governance structures and embedding them into existing systems.
- Dedicating resources to obtain user interfaces that are accessible to and user-friendly for everybody who wants to contribute and actively care for data and knowledge.
- Fostering data literacy in the digital transformation on the three stages: data, information and knowledge.

Of equal importance, if not more so, is the communal implementation of specific measures by all signatories in partnership with Wikimedia Germany. The aim is to promote Wikibase as a promising technical infrastructure for the storage, editing and exchange of data on the basis of the FAIR Data Principles. We wish to shape Wikibase into a user-friendly reference-database software for data hubs in order to promote the desired data ecosystem. To this end, we are inviting further institutions from the world of libraries, from all GLAM (galleries, libraries, archives and museums) areas and from the humanities to use Wikibase in order to create an ecosystem of structured data that comes closer to a true semantic web for FAIR knowledge.[12]

[11] Information on the Fair Data Principles https://www.forschungsdaten.org/index.php/FAIR_data_principles

[12] As an institution, you can co-sign the manifesto via a simple form by following this link: https://www.wikimedia.de/projects/wikilibrary-manifest/

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

The DACH documentation platform[13]

Would it have occurred to you that the article opposite about a football match was written by a computer program? In recent years, the results of computer linguistics have evolved ever further with the aid of artificial intelligence. Writing programs draw content from structured databases and construct the texts with the individual components according to certain specifications. This is the backdrop to our deliberations on recording the frameworks for descriptive and content cataloguing[14] and the data-entry guidelines for the GND as structured data within a Wikibase entity. For decades now, we have been issuing extensive, detailed texts with precise instructions on which data fields must be entered in the GND, for example. Underpinning these texts are the frameworks for descriptive and content cataloguing, the requirements and limitations of the respective software used for cataloguing, and ultimately also the requirements for the exchange of data. Each time a detail is amended at any point in this complex network, the same amendment must also be implemented in many texts that refer to said point. In each instance, this requires labour- and time-intensive research in a large number of PDF pages. Another consequence of this form of knowledge management is that lots of detailed information – such as how to enter a date, for example, or how to record a job description, or which code to use for which country – has to be repeated in various places to avoid having to hunt for said information. When making an amendment, it is important to maintain an overview of every other area impacted by that amendment. There is an inherent risk of errors and a lack of transparency. It certainly makes any guidance less user-friendly, as there is a continuous need for amendments.

The basic principle is strikingly simple. Let us first focus on the GND itself. The number of fields with which one can describe entities for authority data records in the Pica or Marc 21 data formats is manageable at around 300. These data fields or elements serve to make statements regarding

[13] DACH documentation platform: The platform is designed to bring together all the frameworks for library-based cataloguing and the data-entry guidelines for the GND in the German-speaking regions (Germany, Austria and Switzerland).
[14] This refers to the RDA and RSWK frameworks

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

the properties, relationship types, sub-categories or entity codes for the respective entities being described. The data elements contain defined characteristics and different codes, depending on the data format. If all the elements are stored in a corresponding database, the data elements can be assembled in modular fashion, just like a construction kit, according to the rules of the frameworks the database is organised around.

The data-entry guidelines for people alone, with all the requisite entity codes in the GND, encompass 46 pages.[15] Yet the elements to be recorded are few. Along with the name, the other primary elements are the date of birth and death, and any links to other databases, such as place names in the form of the place of birth, the place(s) where the individual worked, or similar. For each of the entity codes in the record-type "persons, a new description is provided each time of how the element "place" must be modelled, for example. If these definitions were to be stored in a database, as a rule, one could simply enter the respective element. This means that if the rule governing the characteristics for recording a regional corporation changes,[16] one can change this centrally in a single location, and all other locations where this element is used are automatically updated too. It is the same principle as applied in the authority data records of library catalogues.

We have started to describe in a structured format all the elements used in the GND. To do so, we are adopting the specifications contained in the frameworks. Now the challenge will consist in writing comprehensible continuous texts in which one can sensibly embed the elements. These can then be updated more concisely than before, and also potentially serve as the foundation for the creation of entry forms for the database with all GND data records.

Sometimes it is beneficial to reflect on the sense and purpose of one's work in order to remain motivated, stay focused and convey to others why this work is important and deserves funding. With this workshop report, we wish to provide you with an insight into our work and the ideas behind it. An exciting time of pioneering work lies ahead of us. This work is made even more interesting thanks to the other, concurrent Wikibase projects in the newly formed consortia of the National Research Data Infrastructure Initiative (NFDI) and other major universal and national libraries in Europe and America, with whom we are in close contact. We will keep you up to date on the latest developments.

[15] Also cf. https://wiki.dnb.de/pages/viewpage.action?pageId=90411361&preview=/90411361/94831186/EH-P-01.pdf

[16] A local corporation is an entity code from the group of geographic entities or places.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12767

References

Harari, Yuval Noah. 2013. *A brief history of humankind*. Munich: Dt. Verlags-Anstalt.

Harari, Yuval Noah. 2020. *Sapiens. The birth of humankind. Graphic novel*. Munich: C.H. Beck.