DEUTSCHE NATIONAL BIBLIOTHEK

BSZ Bibliotheksservice-Zentrum Baden-Württemberg

Same same, but different?[1]

Authority files beyond library use.

Based on the talk held by Barbara Fischer (DNB) and Jens Lill (BSZ) at the CIDOC conference in December 2020

Summary

The digital turn has put numerous museum collections online. This raises questions. How to ensure its unbiased visibility? How to improve retrieval and linkage? Time to take a look at a handy and well established tool in libraries, the authority files. Today authority data files are also increasingly valuable to foster the interaction of scientists working at museums, archives and research institutions. In the course of the digital turn both in science and in culture authority data records fulfil various functions:

- as identifiers and reference for the entity represented
- as junctions within the semantic web where dispersed information on single subjects gets linked
- as a repository for contextual information linked to the identifier enhancing disambiguation of one entity from the other

The Integrated Authority File (Gemeinsame Normdatei, GND) provides about nine million data records for standardized cataloguing mainly used in libraries and supported by the German National Library (Deutsche Nationalbibliothek, DNB). In order to make it a more suitable tool for standardized information for museums, archives and science the DNB has started a five-year research project "GND for cultural data – GND4C"[2] funded by the Deutsche Forschungsgemeinschaft. Together with partners from museums and archives we look for requirements and solutions to open the GND as an information infrastructure for science and culture including its technical, organizational and conceptual properties. The article motivates the wider use of authority files and gives an overview of the results of the first half of the project period.

The authors are **Jens Lill**, working as librarian and museum documentalist at the Bibliotheksservice-Zentrum Baden-Württemberg (BSZ)[3], a state institution operating in South West Germany providing services for libraries, museums, and archives, and **Barbara Fischer**, working as liaison counsel at the DNB[4] in its Office for Library Standards. Both are part of the project team "GND4C".

The article is based on a talk[5] held at the digital CIDOC conference in December 2020. It is addressed to professionals in cultural heritage institutions familiar with the task of cataloguing. It is believed that digital transformation is enhancing their need for linked data. The claim of the article is that authority data may help in the challenge. Thus the article will give an overview of how the project team fosters the participative usage of authority data beyond libraries.

[1] One explanation on the origins of the expression "same same, but different" is found here: https://theculturetrip.com/asia/thailand/articles/same-same-but-different-the-origins-of-thailands-tourist-catchphrase/ (last viewed Jan 2021)
[2] https://wiki.dnb.de/pages/viewpage.action?pageId=134055796 (English version at the bottom of the page)
[3] https://www.bsz-bw.de/ (only available in German)
[4] https://www.dnb.de/us
[5] Online on YouTube: https://www.youtube.com/watch?v=SYE4Szs19Jo

The GLAMorous internet

When talking about the digital transformation, we all tend to think of the internet alone. Albeit we use computers in different gadgets all day long, albeit data has been reduced in its meaning to bits and bytes wherever it is produced or collected and despite that communication in the last months has been mostly in front of screens – still the internet seems to be at the core of all digital transformation. Yet, only taking all this into account and more it sums up to a general digital transformation. Commonly the GLAM[6]-field focuses on the internet being the culminating platform where both content providers and consumers meet and in a more or lesser extent transform into prosumers.

In the last decade most museums have set up their own website. Many have enforced the digitization of their collections. And today many cultural heritage institutions and portals focus on online presence, some more ornate than others. They all follow the idea of showcasing their collections. Some add extensive information to the objects and embed them into a carefully scripted narrative. Some simply opt for a slideshow. They all focus on the curation of content – mainly the highlights of the collection. Very often it is the simple transfer of what museums do in the analogue world into the digital one. The museum exposes its curated collections – be it behind glass or on the screen. The institutional website is considered to be the digital entrance to the museum and its collection. Any London travel guide will mention the Victoria and Albert Museum and indicate its location and website. And if you look for the same museum on the internet you will find it within a blink, no doubt. This is how you will find all its famous artefacts too. Is this a valid

[6] GLAM is a meanwhile widespread acronym for galleries, libraries, archives and museums on the internet.

assumption? Very often a user is more likely to look for a certain object than for the entire collection of a museum. Then the user might use another entrance as s/he might not know to which museum the desired object belongs. It is also possible that s/he is not even looking for a specific object but more its akin.

The impact of Covid 19

public monument masked, credit: public domain by Harvey Boyd via Pixabay

In May 2020 the Network of European Museum Organisations (NEMO) published a highly noted survey on how Covid 19 related measures altered the approach of museums and galleries towards the digital transformation.[7] Confronted with a complete absence of visitors during the first lockdown all participants of the survey reported increased efforts to present themselves and their collections online. There were virtual museum tours, games and competitions involving the digital content, creative social media activities and last but not least more content displayed online. To some extent the lockdown even spurred the cataloguing work as the same persons that normally spend their working hours preparing exhibitions and related matters had now some extra moments to work on their metadata, at least in smaller institutions.

Europeana, the major European cultural heritage aggregator, investigated deeper into it. The debate on the report within the GLAM community as within the Europeana Network Association stressed one thing in particular: The need for further and deeper capacity building inside the institutions, for both handling the hardware and the different technical programmes but also to better understand the nature of data itself.[8]

Harry Verwayen, CEO of the Europeana Foundation, explains this further:

"The reports highlight that the digital divide is much wider than we had previously thought. These divides can be social and technological, as well as between those who can access, are represented, and feel welcomed by digital cultural heritage and those who don’t. The divide also runs between countries who have well articulated digital strategies and infrastructures in place, and those who don’t. They run between institutions that have differing levels of digital capacity and capabilities, and even within institutions where staff have differing levels of digital literacy and skills. Crucially, digital divides are about our processes as much as systems and about people as much as hardware. Bridging these divides will require different strategies, including investigating how our networks and

[7] NEMO Covid19 report: https://www.ne-mo.org/fileadmin/Dateien/public/NEMO_documents/NEMO_COVID19_Report_12.05.2020.pdf (last viewed Dec 2020)

[8] Further details are found here: https://pro.europeana.eu/post/digital-transformation-in-the-time-of-covid-19-edge-predictions (last opened Dec 2020)

narratives become more diverse and inclusive; and in parallel exploring how we can scale up the levels of technological maturity across institutions and countries in Europe."[9]

So it could be stated, the side effects of the pandemic health measures showed their Janus face to many GLAM institutions. The smiling face winked to fully embarque on the digital train with all its opportunities to link to audiences on a global scale meanwhile the other one looked grimmly on the lacking capacity and missing resources to obtain the promised features.

The risk to get lost in cyberspace

Apart from a wide range of capacity levels preventing all institutions to participate on an emancipated level in the digital turn the issue that this article addresses is to improve information retrieval and provide a wider range of tools. Already before the pandemic – as more and more content from museums is available online – mostly addressed through the Google search engine – the single item gets lost in the ocean of data. As habit triumphs an increasing number of users love the Captain Kirk feeling when asking their devices "Ok Google …". Google images was first launched in 2001. Today it accounts for ten billion requests per day. The search engine crawls websites for typical file endings for images, harvests those and stores them as thumbnails on Google owned servers to provide search results faster. As a next step the engine looks for keywords in the title attached to the file or in the surroundings where the file was originally located. This information is used to tag the image. Google claims to display five billion indexed images.[10] No matter typing or asking aloud, the majority of users believe that what they get is a valid representation of what there is to be found. The results may look random to some or are criticized to be the mere commercial interests of Google’s paying clients by others. Anyway, the results are what is taken in account by users. But the unseen rest is deemed to fall into oblivion. Is there a number for the amount of items stored in all museums and galleries worldwide? Most likely not an approved one.[11] But for sure, the number displayed digitally or physically is smaller than what they actually have.[12] And we can all imagine that the number of Google indexed images is only the tip of the iceberg of all images online and the GLAM images among them is only a very small fraction. Yet, despite all the efforts, presenting a selection of the collection online and reaching out, Google still is the mostly used search engine. So what we see is what Google indexes. As many GLAM institutions are fulfilling a public mission, the focus should be to ensure alternative ways to information retrieval.

[9] Quoted from: https://pro.europeana.eu/post/building-digital-capacity-sense-making-findings-and-outcomes (last opened Dec 2020)

[10] Reference link: https://www.bildersuche.org/google-bildersuche/ (last opened Dec 2020)

[11] Some wild guessing on the the amount of paintings is to be found here: https://www.wetcanvas.com/forums/topic/how-many-paintings-exist-in-the-world/#:~:text=All%20it%20means%20is%20that,2%2D3%20for%20every%20person (last viewed Dec 2020)

[12] Reference link: https://qz.com/583354/why-is-so-much-of-the-worlds-great-art-in-storage/ (last viewed Dec 2020). The article is from 2016 and there are likely more art galleries exposing online now, but still not everything they own. As for museums collecting not paintings but artefacts or objects more research needs to be done.

credit: Delft Blue Eyes + Nails, by Francine LeClercq & Ali Soltani, CC0 via the rijksstudio[13]

Sharing the digital content is a promise. All the valuable content that is stored, studied and presented in museums becomes an attractive raw material outside the institution when it is transformed into digital files. The digital presentation of cultural content has an amazing creative potential. In the end, very often the "block busters" become just more known meanwhile less known items are not taken into account. One reason is of course that a meme is only a meme when recognized as such. But more often the reason that content lacks retrieval is simply that it is hard to find. This is the case when providing little or no metadata and linking junctions.

It is not a secret that cataloguing is not the department having the most glamorous appeal. Often a museum’s impact is measured by the number of visitors, be it online or physical.[14] Resources get often allocated based on the parameters of impact. Yet, the impact of cataloguing lacks often a powerful spokesperson. This may be one of the reasons why poor metadata is linked to the items online. Another reason is copyright, at least that is the Europeana experience. Information that is kept in inhouse databases is not shared, but protected by copyright. As the portal claims free licences most institutions opt to hold back their information rather than put it under a free licence and share it. This leads to rather poor description of items on the portal and hurdles the retrieval.[15]

Digital content is not limited to be used in exhibitions, to get reprinted, to serve illustration, to be integrated into memes and gaming. It is not even limited to images. As cultural heritage data becomes more attractive in apps for gaming and edutainment it gains attention for reuse in applications for research too. Science looks at digital content increasingly as big data. New disciplines and professorships in digital humanities have been introduced at all major universities in the last five years, using digital means to analyze huge amounts of data both images and even more text. This leads to new insights.[16] The visualization of data based knowledge has spurred research further. More and more applications for education are on the market based on the content made available through the digitization of the

[13] The rijksstudio is empowered by the Rijksmuseum Amsterdam (Netherlands) https://www.rijksmuseum.nl/en/rijksstudio (last viewed Dec 2020)

[14] Impact beyond audience numbers in this interesting study from 2013: https://online.ibc.regione.emilia-romagna.it/I/libri/pdf/LEM3rd-report-measuring-museum-impacts.pdf and here 2019 on social impact https://www.museumnext.com/article/how-to-measure-the-social-impact-of-museums/ and here https://pro.europeana.eu/page/impact (all last viewed Dec 2020)

[15] Reference link: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/metadata-quality-report.pdf (last viewed Dec 2020)

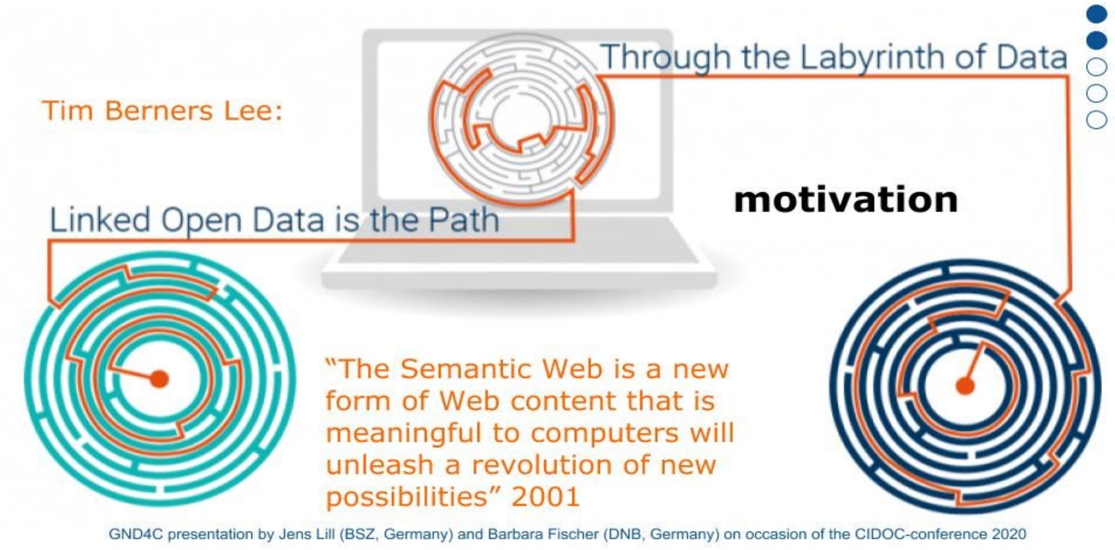
[16] As one example: this article published by the Leibnitz Gemeinschaft (Dec 2020) https://www.leibniz-gemeinschaft.de/en/about-us/whats-new/news/forschungsnachrichten-single/newsdetails/drought-of-the-century-in-the-middle-ages (last viewed Jan 2021)

institutional collections gathered in centuries as our cultural heritage.[17]

Often people in all the fields described above will not come to an institutional website first, when they look for digital GLAM content. They use search engines. As you cannot predict what they will be looking for, the curation of your content on your website will only meet some of the expectations. Yet, the diversity of the items found online is decreasing. The algorithms of Google are trained like a self fulfilling prophecy. The more often a result of a Google search is clicked the higher it gets ranked which leads to more clicks and still higher ranking. The rest is not displayed. Under the subject "painting" Google images lists the following results: After the commercial offers among the first paintings are the ones by famed artists listed like Vincent van Gogh and Leonardo da Vinci and then not much more. What we see is in general from the most popular galleries. Yet, art is just a fraction of the museum field. Artefacts of all sorts form the bulk of our cultural heritage. When looking for traditional clothing in Flanders, Google does not list a single museum or portal as a source for more information on its first two pages. This why we think: Improving retrieval to foster the visibility of the collections and the research performed around it should be an issue to all cultural heritage institutions.

Linked data or the Semantic Web

The most used search engine today was founded 22 years ago. First, we trained the engine. Now, the engine trains our expectations.[18] That is why GLAM institutions too employ masters in search engine optimization (SEO). SEO services aim to improve the keywords used on their clients websites. Another way to spur traffic on the GLAM websites is spending scarce money in Google ads. Is there an alternative?



Through the Labyrinth of Data

Tim Berners Lee:

motivation

Linked Open Data is the Path

“The Semantic Web is a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities” 2001

GND4C presentation by Jens Lill (BSZ, Germany) and Barbara Fischer (DNB, Germany) on occasion of the CIDOC-conference 2020

[17] As an example: the rise and fall of the Roman empire visualized on a map: https://www.openculture.com/2016/10/the-rise-fall-of-the-romans-every-year-shown-in-a-timelapse-map-animation.html (last viewed in Jan 2021)

[18] For the sheer dominance of online research and there Google in specific please read on here: https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/ (last viewed Dec 2020)

Tim Berners Lee's vision of a semantic web[19] is today both more wanted, as more institutions and people share, search and script content in the net, and easier to achieve as the IT capacities have increased. Unfortunately, we have no semantic web up to today. We have mainly one search engine. We are developing artificial intelligence to teach machines to plough through the data oceans faster, more concise and more intelligent. But the semantic web that Lee envisioned almost 20 years ago is not operating. Here is the reason: Most of the data does not speak to machines. As in Wikipedia articles, albeit it is digital content, it takes peta billions of operations for a computer to understand the true meaning of the accumulated data in all these letters. Machines *understand* things by analysing them, by sorting items and by setting up a suitable structure. Text strings are still not the most efficient way for machines to make good use of the data.

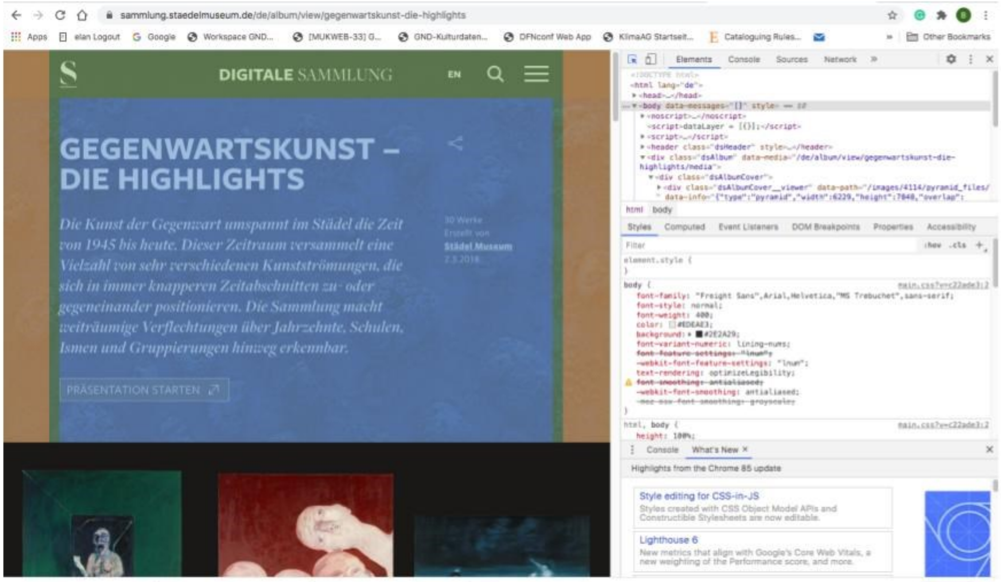
Often the terms "linked data" and "semantic web" are used synonymously. In computing, linked data is structured data which is interlinked with other data so it becomes more useful through semantic queries. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers. Part of the vision of linked data is for the Internet to become a global database.[20]

Imagine all the data would be labeled. Imagine it would be structured and categorized. Imagine there would be persistent identifiers indicating machine readable content highlighting keywords in the content to guide the bot ploughing through the content of all the websites and content displayed on the internet. Then the web content would become meaningful to computers and only then we would obtain the Semantic Web. Being able to process data so much faster the bots would offer us the information needed to answer our questions so much quicker then we can flip catalogues or press keys. And thus offering us alternatives in information retrieval to the Californian search engine.

We would be able to visualize the connections of persons, their hometowns, their skills by a few clicks. Maybe we could show why Mozart lost his impact as star composer and died starving and ill, maybe as a consequence that he started to treat the wrong social circles. We would be able to find among a thousand digitized books from the 18th century all those text parts that would give us a deeper understanding of the *topos* of the noble wild one. Or elaborating an exhibition on climate change, we would be able to trace its impact in art works of the 1960s in the USA. And when researching Calvin's mental framework we would find not only his texts, and items that belonged to him all over the globe, but artefacts related to him and his teachings in places we never thought to look for.

[19] Source of illustration: https://www.ontotext.com/knowledgehub/fundamentals/what-is-the-semantic-web/ and source of quote: https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf (both last viewed in Jan 2021)

[20] Reference of quote: "Linked Data as JSON", https://jsld.org/ (last viewed Dec 2020)



GEGENWARTSKUNST – DIE HIGHLIGHTS

Die Kunst der Gegenwart umspannt im Städel die Zeit von 1945 bis heute. Dieser Zeitraum versammelt eine Vielzahl von sehr verschiedenen Kunstströmungen, die sich in immer knapperen Zeitabschnitten zu- oder gegeneinander positionieren. Die Sammlung macht weiträumige Verflechtungen über Jahrzehnte, Schulen, Ismen und Gruppierungen hinweg erkennbar.

```
<!DOCTYPE html>
<html lang="de">
 ▶<head>…</head>
…▼<body data-messages="[]" style> == $0
   ▶<noscript>…</noscript>
    <script>dataLayer = [{}];</script>
   ▶<script>…</script>
   ▶<header class="dsHeader" style>…</header>
   ▼<div class="dsAlbum" data-media="/de/album/view/gegenwartskunst-die-
   highlights/media">
     ▼<div class="dsAlbumCover">
       ▶<div class="dsAlbumCover__viewer" data-path="/images/4114/pyramid_files/
        " data-info="{"type":"pyramid","width":6229,"height":7848,"overlap":
```

```
element.style {
}
body {                                                        main.css?v=c22ade3:2
   font-family: "Freight Sans",Arial,Helvetica,"MS Trebuchet",sans-serif;
   font-style: normal;
   font-weight: 400;
   color: □#EDEAE3;
   background: ▸ ■#2E2A29;
   font-variant-numeric: lining-nums;
   ~~font-feature-settings: "lnum";~~
   -webkit-font-feature-settings: "lnum";
   text-rendering: optimizeLegibility;
⚠  ~~font-smoothing: antialiased;~~
   -webkit-font-smoothing: antialiased;
   ~~-moz-osx-font-smoothing: grayscale;~~
}
html, body {                                                  main.css?v=c22ade3:2
   height: 100%;
```

Highlights from the Chrome 85 update

Style editing for CSS-in-JS

Styles created with CSS Object Model APIs and Constructible Stylesheets are now editable.

Lighthouse 6

New metrics that align with Google's Core Web Vitals, a new weighting of the Performance score, and more.

The illustration above shows the code behind the online presentation of a famous gallery in Frankfurt.[21] The online exhibition is based on data from their database. The code online is readable and accessible to machines. The more structured data we include the easier it becomes to find data even though you do not know where to look for it. But the data is there and can be found. That is the basic idea of retrieval.

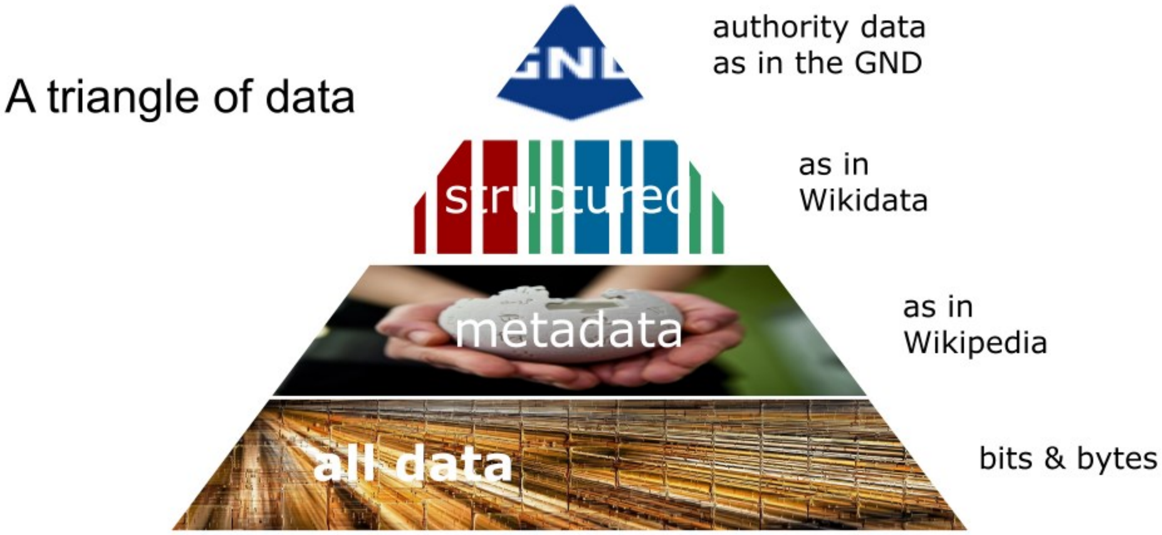
It needs cataloguers

The Semantic Web does not come automatically. It is not based on some cryptic algorithms or artificial intelligence. They will help the cause. The simple truth is, it will be based on human labor. It needs people. People that care for cataloguing and entering metadata into their databases and sharing it. As the amount of data is increasing enormously the more relevant the issue of improved retrieval and visibility of the displayed and stored content gets. The requirement to ease re-use of data becomes crucial to minimize the risk to fall into the digital oblivion. The value of the data increases by being used in more and diverse ways.

Often we limit ourselves asking: What is needed to attract spectators, to engage with the visitors? How to turn them into members of the museum community? As civic identity is rooted in our cultural heritage, its mediation through museums fulfills a democratic mission. So museum strategies raise the stakes as does the entire GLAM field. Now, we learn this might not be enough. When the numbers of visitors drop. When the pandemic – like under a burning glass – reveals our vulnerability in changing times. What is the complement to the visitor? We need to be able to cope with the new requirements. Online is no longer behind the screen like we used to display our collection behind the glass of our vitrines. Our audience has moved. Art and creative industries have moved. Science has moved. We have moved. We are no longer in front of the display cabinet. We are all increasingly inside the digital world. The GLAM employee needs to become a data wrangler. We need to obtain

[21] The code behind an online presentation: Digital collection of the Städel Museum https://sammlung.staedelmuseum.de/de/album/view/gegenwartskunst-die-highlights (last opened Dec 2021)

data literacy. Treating, handling and serving our digitized collections as data, actually as big data. All data is the same basic electric impulse. But some data is different.



This simplifying triangle explains the differences within data. At the bottom there is all data. No matter whether images, 3D representations, film, sound, text, letters, numbers … all is reduced to either the electric impulse is on or off. On the internet the Nefertiti is just a bunch of zeros and ones. At the next level the data is transformed into metadata. The articles in the encyclopedia Wikipedia describe the items. It is digital data but still not machine readable. The third level is the structured data as in Wikidata. This data can be read by computers. Wikidata became necessary to maintain the quality of the Wikipedia articles keeping them up to date through centralization. In Wikidata the number of inhabitants of Paris is updated at one place and gets it changed in all Wikipedia articles no matter if in Urdu or in English. In Wikidata an item is described by properties with their defined qualifiers. It has a label and a persistent number starting with Q. Paris, the French capital, is represented by Q90. Yet Wikidata itself relies on authority data. In order to be able to reliably distinguish one Paris from another one in Texas (Q830149) and to link the information found on the French capital to its Wikidata identifier, the Wikidata community links it to approved sources and the identifiers they use. These identifiers are provided by libraries across the world in their authority files. The GND is this Integrated Authority File to the German-speaking countries.

To claim authority

The GND is held by the German National Library and provided under a public domain licence. Everybody is entitled to use their ID-numbers. Many GLAM institutions do so. The German Digital Library (Deutsche Digitale Bibliothek, DDB), the national aggregator to the Europeana cultural heritage portal, recommends the use of GND-IDs and whenever some extra time is available they enrich the metadata on the items exposed with GND-IDs before handing them over to Europeana. The GND covers all sorts of topics, of persons and sites. So why do we need an integrated approach across domains, disciplines and sections? The simple answer to that question is: because the work needs to be done by somebody. Authority data is a handy tool. Yet, who carves the tool? To create authority files, to maintain them accurately and to ensure their quality takes effort. It is nothing that happens automatically. Librarians create authority data records but they create only the records they need when doing subject cataloguing.

When thinking of central activities in libraries these might come in mind, amongst others:

- Cataloguing and sorting books or other media
- Making sure you will find documents and their content
- Grouping and classifying
- Managing information
- Organizing knowledge

Maybe the very core of the library profession is the last one: Organizing knowledge. It has always been. However organizing knowledge has become more attractive to other communities due to the general digital turn. As more data is available online it is more difficult to find exactly the piece of information you are looking for. Here people either use search engines or – if looking for deeper results – they turn to the library of their confidence.

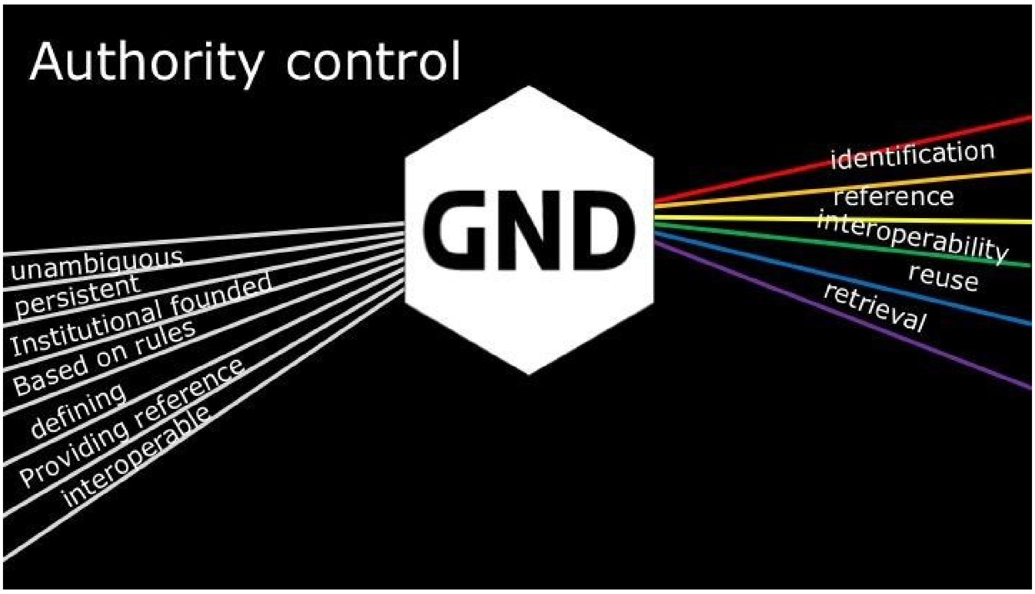
The basic idea

One of the central tools organizing knowledge in libraries are authority files. They link one identifier – normally a numeric code – to a label or a name. This means they codify a text string which makes it readable both to humans and machines. By adding a persistent link you get a reference point in the web. That is the basic idea:

label - ID - URL

But this alone is not enough! It would only be a list of names, numbers and URLs. There is one important thing missing. Disambiguation. Take the personal name Alexandre Dumas. Most people know that they have to distinguish the elder who wrote "The Count of Monte Cristo" or "The Three Musketeers" from his son Alexandre Dumas, who wrote "The Lady of the Camellias". But probably only few people know the Canadian historian with the same name. A name or label is often not enough to clearly identify an entity. You need a further property related to the item you like to codify. As to persons, disambiguation often is achieved by relating it to their birth date as identifying property. For places the georeference is suitable. Works are linked to their creators and so forth. Being very much down to the basic rules, the formula would thus be:

label ↔ ID ↔ URL + at least 1 disambiguating property



An authority record is expected to deliver identifiers, persistent URLs and reliable identification of the described entities. Meeting these requirements, the authority file GND is that tool for the German-speaking countries. The GND provides IDs to categories such as persons, geographic names, subjects, corporate bodies, conferences and works being referred to in publications either printed, digital or in other media.

The GND uses an Entity-Relationship-Model (ERM) and has a modular data structure. The entities have properties and are related to other entities. Relations are defined by codes. The GND is based on international and national rules or standards. It serves standard data formats such as MARC21 and RDF in different serialisations. The GND converged in 2012 from four precursors and contains today about 9 million records. Thus its name: Integrated Authority File or – in German – Gemeinsame Normdatei. It is maintained by the German National Library, member of the GND cooperative, and is provided under a public domain license for free reuse (CC0). The GND wants to further the retrieval, identification, contextualisation and attribution of data items.

These are the hard facts. But there is what may be called a soft feature too. The core quality of an authority file is its reliability. People using the GND have faith in its institutional contributors delivering high quality data. Yet until now the GND is a quality-tool produced by librarians to serve mainly their needs.

Things are changing

Now we experience the impact of the digital turn. In the past years authority control has gained a certain appeal among museums, archives and research in general. There is an apparent need for persistent identifiers. What distinguishes the GND from other library authority files is that it is already a product of cooperation. Approximately a 1000 libraries in German-speaking countries – Germany, Austria and Switzerland – contribute to the GND, and a lot of more use it. Wikipedians do not only link Wikipedia articles to it but help to correct and enrich entries especially on persons. Therefore its name "Gemeinsame Normdatei" could also be translated to Common Authority File.

In the past years, the administrative structure and strategic programme of the GND have been changed little by little. The goal is to enable more interaction, more cooperation, more integration of data, which is not conceptually linked to publications, reaching out to communities beyond the realm of libraries thus increasing both the amount of items represented in the GND but even more their diversity.

In 2017, twelve partner associations have signed a contract that regulates their rights and duties in the GND-cooperative. It defines workflows and standards to the quality of data delivered. It is the first legal framework of the GND.network[22] and it enables new partners outside of library structures to join the cooperative.

To open the GND

The research project "GND for cultural data" (in short: GND4C) started in May 2018. It is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). Five partner institutions from different realms are investigating what it takes to open the GND for non-library use. Experts from libraries, museums and archives define the specific requirements of the latter on the organisation, the data model and the technical infrastructure. The goal is to enable not only the use of existing GND entities in their own databases but to add new data records to the GND and update existing records on their own behalf. We are looking both into the theoretical framework and prototyping practical workflows. At the same time we are evaluating the potentials of Wikibase as a new software base for the GND operating alongside the existing library IT-systems. The project investigates four fields and strives for four goals:

- data model and rules – striving for consensual concepts for non-library use
- infrastructure – providing APIs and applications for non-library requirements
- governance – building a cross domain organisation
- communication – increasing the visibility of the GND.network

Defining the necessary and indicating the diversity

Authority control builds on rules and standards. Librarians have been trained to achieve the maximum quality following established standards for generations. There are definitions and settings that have been long debated before settled and integrated into an international framework where all these procedures are reflected as well. As a result, it is quite a challenge to reach out to other communities that do not share the same framework. We have only started to compare our data model with those in use in museums or archives. And there is not one other data model but many. But for now most cross domain requirements concerning persons, geographical names, corporate bodies and events as conferences are fully met by the existing GND data model. Yet, general concepts or subject headings and the dazzling character of the concept "work" need further research. In general the GND data model is flexible enough to cope with the requirements of museums and archives. The

[22] For further information see https://gnd.network/Webs/gnd/EN/Home/home_node.html (last viewed in Mar 2021)

research by the project has proven that there is a need to give a greater independence to the regulations treating the authority file itself from the cataloguing standards. Here we suggest to install a core rule set and allow additional community based rules, so called plus-sections of regulations. Further we aim to codify the affiliation of the data records to make them readable both to machines and to humans. The idea is to indicate the parts of the record that are relevant to all and fulfill the core needs of identifying and defining qualities of an authority file. Those properties or their community specific qualifiers that are limited to a specific community but not binding to others will be marked as plus-section. For one community the profession of a person needs to be gendered as others refrain from doing so and relate the person to the record of that profession within the GND.

Providing tools for automatic data reconciliation

In order to make it easier for other communities to integrate their data to the GND and to avoid the deterioration of the quality by creating duplicates through mass import of data we are creating tools to help the data check. It is a big challenge to enable automatic solutions exchanging data from and to multiple systems because not only data exchange formats differ but also which entities are catalogued and in what way. There are many software providers and probably even more data models and cataloguing traditions. There is no general standard how to describe an entity be it an object, a concept or a person as there is no euro plug that fits all the sockets types in Europe. The following illustrates the problem. In order to be able to check if all contributors to a museum collection have a GND-ID, their names need to be extracted as entities from the museum’s database. The name alone, as shown above, is not enough. A disambiguating property is needed as well. All the information needs to be consistent and structured data in order to be processed by the computer program. Very often the information is contaminated by unstandardized textstrings. The data has to be refined beforehand. In order to improve the reliability of the results the provided data is compared both to the GND and other databases such as Wikidata or GeoNames. This means a lot of data processing has been done before the actual matching with the GND-data records can start. The diversity of the cataloguing data hampers the establishment of generic workflows and algorithms. The goal is to develop a tool set that will sort the data on various entity types into three categories:

- Match, meaning the applicant data record matches fully its equivalent GND record. Import is denied. The GND-ID is reported back to the data provider.
- Partly Match, referring to a data record that is likely to match a GND record, but shows differences that need to be checked to make sure, if it is a match. Under the established ruleset the GND may import extra properties from the applicant data record. After approval: The GND-ID is reported back to the data provider.
- No Match, the applicant record is unknown to the GND and fulfills its criteria for eligibility. A new GND record will be created and the new ID number is reported to the applicant.

Ideally the toolset will be used autonomously by GND agencies on behalf of their clients in order to improve their data quality. For now, this is left to be constructed.

Enabling a cross domain governance structure

In the German-speaking countries the institutional landscape of museums and archives is far less integrated into a national and international organisational system than the section of libraries. In the field of museums, many institutions are working on their own and without adequate financial or human resources. The section of archives is rather segregated too. The project GND4C tries to find solutions to the fact that there is neither one representative of all archives nor of all museums in order to agree to standards on how to define the core properties of the GND records. Instead we invite communities who are interested in the matter to organize themselves as interest groups and name a spokesperson in the Committee of Standardisation.This Committee decides on the standards applied in cataloguing for the German-speaking library community and subsequently for GND Affairs, too. This indicates both a vast and a complex agenda. The expertise is very specific to each community and there has been little need for exchange across the boundaries of domains, until now. We are at a crossway. Is there a need strong enough for a cross domain standardisation in cataloguing? Or should we rather limit the efforts strictly to the GND itself? The second half of the project will tell.

Forming agencies to reach out for more diversity

As one result of the project we have established the idea of GND agencies. In December 2020 the first agency started addressing museums and archives in the state of Baden-Württemberg. Their role is to consult and inform institutions interested in the use of the GND. They will operate as editorial departments if new GND records are required. Only those records will be admitted that meet the eligibility criteria. These criteria have been established recently motivated by the project.[23] Until now, the consent on the criteria was kept informal. If there is a publication stating a new author, a new subject or a new corporate body, then a new GND record is necessary. Confronted with the possibility of mass imports of new records from museums or archives the eligibility needed to be formalized. It is still ruled by need but that need has to come along with reliability both concerning the data and the institution willing to take care of the record, its quality and transformation in the future. Only if an institution is ready to take that responsibility or delegates it to a GND agency, the mass import of new records into the GND takes place. The agencies will perform the data reconciliation on behalf of their clients and train them in GND editing. A second agency addressing monuments and protected real estate will soon start to operate. For now the agencies are mainly funded through the project itself. The challenge will be to shape sustainable business models. There is a request for further agencies. The regional approach of the established library service centers acting as GND agencies serving mainly libraries is reflecting the federal constitution of its member states. If it can or should be transferred to museums, archives and further research centres has to be proven.

The project GND4C and everyone in the opening process at large is focused on the requirements of the new "customers". But what about the library world? Does it have to change? Does it *want* to change? We have to advocate the necessary changes in all fields

[23] They are published in German here under the heading "Eignungskriterien" https://gnd.network/Webs/gnd/DE/UeberGND/GNDEignungskriterien/eignungskriterien_node.html (last viewed in Mar 2021)

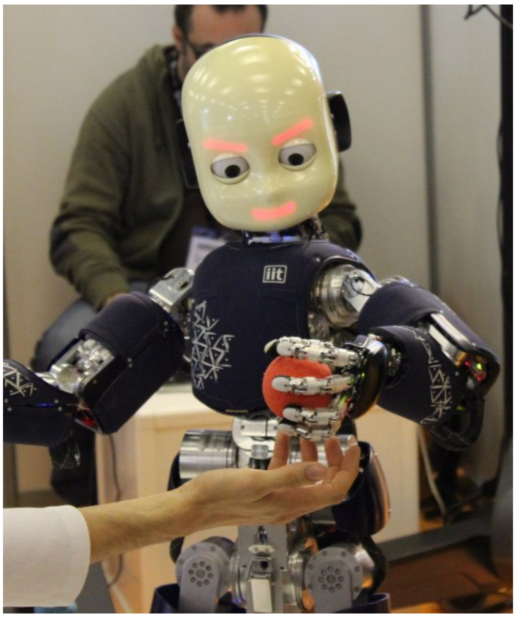
in quite a traditional community – and that is time consuming. For now we can only state, all that change requires efforts and patience from all stakeholders. The interest shown from many institutions to make use of the GND is boosting our motivation to strive for solutions.

Instead of a summary, a call

To sum up, why should you care? You are cataloguing your collections and presenting the data on the net to a wider or lesser extent. You are looking for additional content to your own data and you know others might consider your information as interesting data for themselves. Exchanging data has become the crucial requirement of our time. But linked data is not enough. The Semantic Web is more promising.

The Semantic Web gives a structure to the ephemeral data – in the digital, structure is the key. Who will be able to understand our digitally born thoughts in 50 years, when software and hardware change as much as they have in the past 30 years? We need to ensure the linkage to its context through reliable and persistent data as meaningful information. We need a flexible, yet stable response to a rapidly changing digital environment. We need a semantic web we can refer to and bind data to.

Scientific thinking could be described like: We take an object apart, we look for rules and relations, we define properties and we fit it into a system of categories that becomes a new object. We structure the data to obtain information and to transform it into knowledge. The workflow we apply is similar to algorithms used by computers. But structured data is not a machine thing. To structure data is a human behaviour we teach machines to copy.



© Xavier Caré / Wikimedia Commons / CC BY-SA 4.0
https://commons.wikimedia.org/wiki/File:ICub_Innorobo_Lyon_2014.JPG

Authority files support the structuring of data. Shared persistent standards help communication between humans and links to coded identifiers it helps communication with machines too. It increases the velocity of retrieval and exchange. This helps the evaluation of ideas and the analysis of information. It opens the perspective to more knowledge. A better understanding of the task ahead.

Librarians knew all that even before transforming authority files into a digital tool. Museums, archives and science institutions in German-speaking countries aim to share that handy tool as they see their data online getting lost in myriads of entries on the world wide web. This is why they put resources into becoming part of the GND.network.

They want to improve retrieval by linking their entities to GND-IDs, rationalizing their cataloguing work by importing data from the GND-records and increasing the diversity and richness of the GND by adding new records to it.

This article ends with a call. We would like to know, where do you turn to obtain reliable structured data? Do you have access to the authority files of your library colleagues? Which entity types do they cover and how suitable are they to you? Are you participating in the creation of new authority records satisfying your needs? We would like to link our project GND4C to similar projects in your countries. So do not hesitate to reach out to us and tell us how cataloguing and authority files spur the Semantic Web in your environment or if you opt for a completely different approach. We are curious.

(Text: March 2021)

CIDOC 2020 in Geneva, Switzerland
hosted online by the Museum of Art and History of the City of Geneva

Session 6, Connecting object information with archival and library resources
Tuesday, 2020-12-08

Paper 4: Same same, but different? Challenges and solutions in the opening process of the GND authority control for cultural institutions

Authors

Barbara K. Fischer (GND-ID: 1242064478)
German National Library (DNB), Office for Library Standards
Deutscher Platz 1
04103 Leipzig
E-Mail: b.k.fischer(at)dnb.de

Jens M. Lill (GND-ID: 1059591952)
Library Service Centre Baden-Wuerttemberg (BSZ), Museum Information System
University of Konstanz
78457 Konstanz
E-Mail: jens.lill(at)bsz-bw.de