Ongoing efforts to build a robust infrastructure for collections (including books, images, etc.).

The goal is to optimise for speed, e.g. by using Elasticsearch as a foundation of the infrastructure.

Enrichment of bibliographic records from different sources such as bibliographic records from other library networks by clustering work bundles or ORCiD. Links to authority records for persons and elements of content description are transferred from other cluster members. Matching processes allow enrichments of person authority records with unique identifiers such as ORCiD and ISNI.

Automatic subject indexing with Annif. Automatically assign GND descriptors, DDC Subject Categories, DDC Short Numbers Visualization of authority file records (Integrated Authortity File (GND) and statistical details.

The German Museum of Books and Writing uses their collections and DNB data for text mining (Analyzing the content of digitized and OCRed holdings) and Image processing (Tests with AI methods for analyzing watermarks).

The library automation team within the University of Antwerp Library maintains a library system Brocade which is implemented in multiple research and special libraries and archives in Antwerp and Limburg provinces, Belgium. In addition, Brocade is used as platform for an institutional repository in multiple institutions, including the University of Antwerp, and for a database for research output in the social sciences and humanities in Flanders. Hence the data available in our system is rather broad and include bibliographic metadata, full text pdf's, digital material, and research data. The library system hosts several heritage collections (i.e. collections of historical manuscripts and works of art).

At the moment, bibliographic metadata from the Anet library network and other projects maintained by the Brocade library system can be made available in the following formats: sqlite, csv, pdf, xhtml, html, tag, xlsx, xml. Some of the data are published as open data sets in XML format on the Anet webpage: https://www.uantwerpen.be/nl/projecten/anet/open-data/ In addition, catalogue data are accesible through Z39.50 and OAI-PMH.

Digital material from the University of Antwerp Library collection ‘Preciosa’ (items published before 1830) is available on digital platform: https://anet.be/query/uantwerpen/opacuantwerpen/ua-dp

At the moment, the library automation team is implementing IIIF framework that would allow a more flexible interaction with images within the collection. To make this possible, existing images were converted in order to comply with the IIIF Image API. A manifest generator was dAPIeveloped allowing the distribution of images and corresponding metadata in compliance with the IIIF Presentation.

Concerning metadata enrichment, the library automation team has implemented automated language detection for the titles of bibliographic records. For the institutional repository, there’s an automated retrieval of citation information from Web of Science. There’s also a semi-automated upload of Web of Science data to the institutional repository and as part of this process author authority records are enriched with ORCID. Also, there’s a tool in the institutional repository that allows researchers to submit their publications by entering a DOI or pubmed identifier. On the basis of these identifiers, metadata are automatically retrieved from Crossref and Pubmed and afterwards checked for quality by catalogers.

Data from the library system are supplied to third party service providers that include the Union catalogoe of Belgian University and Special libraries (Unicat.be), OCLC WorldCat (worldcat.org), Ebsco Discovery Service. Institutional repository of the University of Antwerp is linked with ORCID whereby researchers from can update their publication lists in ORCID using the data from the repository.

UN SDGs mining project: Identification of UN sustainability goals in research publications (R&D Department at SUBn Göttingen):

In 2015, the United Nations Member States (UN) agreed on 17 sustainable development goals (SDG) „for peace and prosperity for people and the planet, now and into the future“.

An SUB research group wants to support this endeavour by clustering research publications, based on their relevance to these SDGs. The goal is to make relevant research findable and also to strengthen access and visibility to such research.

A combination of approaches is necessary to deal with diverse challenges. First, texts must be prepared for processing. Since research publications are written in different formats, styles, semantics and even languages, this is where the first challenge begins. Policies must also be considered. A variety of licenses and conditions have to be considered. Once this is done, the content of the texts can be analyzed. A variety of methods are available for this purpose: A combination of text and data mining, topic modelling and linguistic analysis is necessary to cope with the challenges of heterogeneity of texts, different semantics and large data sets.

The group aims to compare several methods applied to a corpus that is ready for take-off.

TextAPI (https://textapi.com/): putting the power of natural language processing into the hands of the everyday computer programmer or industrious business owner. It is surfacing this power to do text analysis, extraction, and manipulation via a language agnostic web API. This allows a developer of any walk of life whether it be PHP or C# to be able to leverage this potential.

We are setting up Annif (www.annif.org), a toolkit developed by the Finnish National Library, to automate subject indexing and classification. In order for the algorithms implemented in Annif to index and classify future documents, they need to be trained on previously indexed ones. We are working on collecting this data from existing bibliographical databases and digital collections.

One current data science project is the (semi)automatic deselection of books. Deselected items are flagged for physical removal. The automisation step makes use of the OCLC API to retrieve holding counts from other Dutch universities and other item properties. This information is used to recommend items for the deselection process.

For a few years an electronic tool is used to collect data on internal workflows in the library, e.g. indexing (kind, depth), number of incoming legal deposits (print, online), duration of processing incoming media.

Statistics of media orders in the reading room relating to time of orders and type of ordered medium were maintained manually until recently and frequency of visits and orders are analyzed as well.

These data are published in the Annual Report. They also serve as a basis for internal planning e.g. concerning the allocation of work and staff, quality control, provision of support for selected tasks in the process.
Item suggestions in DNB´s catalogue are implemented via a commercial tool analyzing user behavior in the catalogue (bibtip).

The different departments within the library regularly produce reports and statistics both for internal use and for external organisations (primarily, regional or national). There's also ongoing monitoring of the use of library resources. Data sources for this include the library system, publishers, database providers. At the moment, this work is carried out using a combination of tools within the library system and MS Excel. Some staff members export data from the library system as csv or sqlite and then continue their work with R or Python.

Since the library system is developed by the library automation team at the University of Antwerp, occasionally the library automation has to carry out some data engineering tasks in order to provide the data that the library staff requires.

Statistics about the proportion of OA publications in order to influence OA policies and collections management.

SSHOC (Social Sciences and Humanities for the European Open Science Cloud, https://sshopencloud.eu/, an EU-funded project): building a publication network pipeline as addition to an existing metadata extraction framework

Digital Geochemistry Infrastructure (DIGIS, https://www.uni-goettingen.de/de/digital+geochemistry+infrastructure/643369.html): extracting information from research papers for a geochemical database

Culture Cloud: detection of cross reference between objects in collections at campus Göttingen (prototype for a proposal).

OCR-D (https://ocr-d.de/) projects (OCR-D Coordination, OLAHD, OPERANDI) to provide the requirements for data science activities on digitized books.

Text+ (https://www.text-plus.org/en/home/), national research data infrastructure for language- and text-based research: several different TDM related tasks

QA catalogue: quality assessment of MARC21 library catalogue to analyse metadata records and point out issues. Used in 3 libraries (U Gent, British Library, Belgian national library) to run an analysis on a regular basis (e.g. weekly) and fix the issues (https://github.com/pkiraly/metadata-qa-marc/)

Metadata Quality Assessment Framework API: a genric tool to assess metadata quality. It is used in Europeana, German Digital Library, Victoria and Albert Museum, meemoo (Flemish audiovisual archive) (https://github.com/pkiraly/metadata-qa-api/)

Research data management training and support; software carpentry training for researchers

Collaborations with researchers in individual projects where DNB provides selected metadata content or access to full texts on the premises (e.g. Children and Youth literature, dime novels)

DNBLab is targeted at Digital Humanities researchers. It offers several collections of freely available data, both metadata and full texts. It also offers tutorials and webinars to teach data analysis techniques.

DNB offers a yearly DH call. Applicants can obtain access to DNB data and infrastructures to work on Digital Humanities projects focusing on DNB content.

DH Fellowship: investigate DNB´s metadata and digital content for research purposes and receive a funding for a period up to 12 months

Participation in NFDI4Culture a consortium funded in the German National Research Data Infrastructure (NFDI) framework concentrating on cultural heritage.

Dataverse built-in tools (e.g. Data Explorer (https://scholarsportal.github.io/Dataverse-Data-Explorer/))

Research group specific data repositories: e.g. TextGrid (Digital Humanities), FOR 2432 (agriculture), Bexis (biodiversity), DIGIS (geochemistry)

GeoMapper (https://geomapper.goettingen-research-online.de/, repository: https://gitlab.gwdg.de/era-public/dataverse-mapper/), developed by the SUB R&D department

Our institution, in collaboration with the Department of Scientific Computing and Research Support (DCSR) at the University of Lausanne, has launched the Machine Learning Club. It serves as a place to discuss and share experiences on topics related to machine learning and artificial intelligence.

We conduct analyses for internal and external clients. We deliver dashboards and reports that are usually tailored to the specific needs of the client which tend to vary between projects. The activities therefore include harvesting and processing the data using Python, data modeling and visualisation. We use a variety of data sources such as Scopus, Pubmed, Altmetric, Unpaywall, etc.

Two main activities: compiling and visualizing publication data for (1) the national consortium negotiating with scholarly publishers and (2) the national evaluation of the transformation to open science, in particular open access to scholarly publications.

Analysis of big scholarly data, reports and visualisation (e.g. interactive reports, dashboards):

* Open Access Uptake in Germany https://subugoe.github.io/oauni/
* Measuring Hybrid Open Access Uptake using open scholarly data sources (in progress, Data R package https://subugoe.github.io/hoaddata/)
* metacheck, a self-assessment tool for open schoalrly data build implemented as a R package https://subugoe.github.io/metacheck/
* Scholarly Communication Analytics blog, https://subugoe.github.io/scholcomm_analytics/

Skills for working with data - targeting both librarians and researchers

* Carpentries working group (https://pad.gwdg.de/GOECarpentries#), acrosss various SUB units, GWDG and university faculty; contribution the Carpentries community (e.g. member in Library Carpentry Advisory Group; maintainer role for LC lesson)
* Book club on Data Science
* Short course on bibliometrics / digital science at Humboldt University for post-graduate LIS students ([Slides](https://docs.google.com/presentation/d/1D-lEnMtgKSqhQUZzxgAa3kiQaYW8H0rVOkbZkO7VYmw/edit#slide=id.p))
* Session on reproducible research at Göttingen Open Science Meet-up, Oct 2020

Contribution to a study on publishing reproducible research outcomes (via Knowledge Exchange Open Access Expert Group, https://doi.org/10.5281/zenodo.5521077)

- ON-MERRIT (EU project): Patent analysis on open access in innovation

Research publications since 2020, e.g.

* Jahn, N., Matthias, L., & Laakso, M. (2022). Toward transparency of hybrid open access through publisher-provided metadata: An article-level study of Elsevier. *Journal of the Association for Information Science and Technology*, 73(1), 104–118. https://doi.org/10.1002/asi.24549

* Fraser, N., Hobert, A., Jahn, N., Mayr, P., & Peters, I. (2021). No Deal: Investigating the Influence of Restricted Access to Elsevier Journals on German Researchers' Publishing and Citing Behaviours. ArXiv:2105.12078 [Cs]. http://arxiv.org/abs/2105.12078

* Hobert, A., Jahn, N., Mayr, P., Schmidt, B., & Taubert, N. (2021). Open access uptake in Germany 2010–2018: adoption in a diverse research landscape. *Scientometrics*, 126(12), 9751–9777. https://doi.org/10.1007/s11192-021-04002-0

* Laakso, M., Matthias, L., & Jahn, N. (2021). Open is not forever: A study of vanished open access journals. *Journal of the Association for Information Science and Technology*, 72(9), 1099–1112. https://doi.org/10.1002/asi.24460