TRAIL 5.3:

Sustainable retrodigitisation of paper-based research data

Partner **Lead:** General Directorate of the archives of the Bavarian State (GDA: Schmalzl)

Co-applicants: Verbundzentrale des GBV (VZG: Dührkohp)

Participants: Archaeological Museum Hamburg (AMH: Räther), Thüringische Universitäts- und Landesbibliothek (Thulb: Christoph), Niedersächsisches Landesamt für Denkmalpflege (NLD: Statje), Centre for Baltic and Scandinavian Archaeology (ZBSA: Schmölcke), Niedersächsische Staat- und Universitätsbibliothek Göttingen (SUB: Rühle); Heidelberg Center for Cultural Heritage Heidelberg University (HCCH: Börner, Sieckmeyer)

External members: Text+, NFDI4memory

Contact Markus Schmalzl / Poststelle@gda.bayern.de

Summary

At numerous institutions in and outside the community, extensive scientifically valuable research data is insufficiently indexed and only accessible on site, in paper files, official records, card indexes, maps and plans, photographs and other analogue data storage media. These data can only be used to plan field research or object-related studies or evaluated in metastudies etc. at enormous expense. For example, the analogue files of the lower heritage authorities and their official predecessors on archaeological finds dating from 1820–1970 (which are stored in various archive holdings at locations across Bavaria) can be made available for joint analysis online through retrodigitisation and enrichment with sufficient metadata. The challenge here is to develop a modular tool-based process that N4O can offer as a service. This includes excavating files, digitisation, indexing, OCR indexing if necessary and transfer to long-term archiving (LTA). It takes into account legal and professional restrictions on use (data protection, copyright, sensitive data, e.g. to localise cultural assets) as well as

technical and professional standards. It thus covers the whole research data lifecycle: publish, share, preserve, reuse, create, assure and describe. The TRAIL will create a core retrodigitisation service for N4O, using standards data and providing the metadata in the standardised exchange format EAD. It will also produce a publication of the holdings for subsequent use in research and teaching and a concept for LZA.

Description

Many older research data (files, excavation documentation), which are highly relevant for current research, are only available in paper form, with insufficient metadata and across different institutions. This makes them difficult to research and usually only evaluable at great expense. Automated evaluation is unthinkable without preparatory work. The TRAIL will develop a solution according to M5.2 of TA5. Metadata sets for analogue document groups are developed in coordination with TA1, TA2 and the requirements of the discovery system to be developed in TA5, LTA, and legal or subject-specific restrictions on use. In this TRAIL, we will retrodigitise collection documentation of the numismatic collection of Heidelberg University (approx. 300 pages); paper record and measurement lists of archaeological and zoological finds of the Centre for Baltic and Scandinavian Archaeology and predecessors in 1965–1990 (approx. 90 files / 18,000 pages); files of the lower heritage authorities and their predecessor institutions in Bavaria on archaeological finds in approx. 1820–1970 (approx. 120 files / 16,000 pages). Research materials will be prepared and scanned based on the specialist digitisation concept of the Bavarian State Archives. Based on Goobi software, a workflow for the digitisation and publication of archive holdings is being developed, which will be provided as a core service for N4O. A suitable export interface will be created to prepare the data and metadata for partially automated archiving via the generalised XML client of the Bavarian State Archives. The specialist digitisation concept is based on the Bavarian State Archives' decades of experience in archival digitisation. With this experience, existing standards and authority data were improved and incorporated into a new version of the digitisation concept in 25 July 2019. Cultural change in archiving over the past decade is now being supplemented by community-driven processes. The XML client of the Bavarian State Archives relies on automated ingestion, structuring and acknowledgement of data. All processing steps are documented to ensure data provenance, using the latest technology.

Commons

The goal of the TRAIL is to create a core service of N4O. The metadata models will be published as a white paper and the workflow as a blue paper to be discussed with the community, a process guided in TA7.

Relevance

Up to now, extensive databases on archaeological excavations, object finds, provenance and object histories have had limited availability to researchers, as they are insufficiently indexed, only held on analogue data storage media and decentralised. This is where the TRAIL comes in, addressing the publish, share, preserve and reuse aspects of the research data lifecycle.

When the metadata and the digitised material are put online, researchers will have greatly improved access to data that was previously virtually inaccessible. Data curators and infrastructure providers will have a tool-supported workflow based on professional standards and thus a core service to retrodigitise analogue research data. The data is available to researchers via standardised exchange formats and interfaces for integration into information infrastructures for subsequent use and is permanently secured via LTA.

Other non-participating sub-communities in N4O benefit just as much from the standards for indexing and metadata of (excavation) documentation and files. For LTA, the data is converted into a form that can be processed automatically, enabling subsequent evaluation and networking with data records from other sources. Existing standards are developed and made available for users to apply to their own systems. In addition, data interfaces are provided for exchange.

The service provides data in the international exchange format EAD, the datasets are largely standardised through the consistent use of authority files on the vocabulary service DANTE and LTA standards. The experience gained in this TRAIL will be applied to other analogue document groups relevant to the professional community.

The service meets all the FAIR criteria. Enrichment of existing metadata, retrodigitisation and reference in the N4O discovery system make the data findable, accessible and interoperable, and reusable through LTA in IANUS, the Digital Archive of the Bavarian State Archives or another public repository of N4O. The developed software environment can be used within the consortium and beyond; the prototype process can be adapted to alternative software solutions. To date, no core retrodigitisation service has been available for the community or for the NFDI as a whole. In addition, addressing legal issues guarantees the subsequent use of sensitive research data. Networking to develop standards and exchange formats is planned with Text+ and with NFDI4Culture; with the latter, we will also address legal issues (retrodigitisation guidelines). FAIRification of analogue research data is relevant to other NFDI consortia (4Earth, 4Biodiversity, KonsortSWD).

Deliverables

The development is based on the open source software Goobi and the generalised XML client and specialist digitisation concept of the Bavarian State Archives. Optional interfaces for additional services (e.g. OCR, LTA) are available. The retrodigitised material and its metadata are made available via a standard viewer in METS/MODS and Dublin Core formats via an OAI-PMH interface. By mapping the tectonics in EAD, it can be integrated into archival reference portals (e.g. Archive Portal D). The digital copies are available via an IIIF interface. The specialist vocabularies are integrated via the DANTE authority file service.

Existing data formats and interface descriptions are developed and documented and are available for use in other systems. The core service ensures the permanent availability of archive holdings and previously paper-based primary data for research.

The data and software environment will be used in the curriculum of the MA degree programme Cultural Heritage and Protection of Cultural Property at Heidelberg University and on the e-learning platform NumiScience.de, thus ensuring transfer into teaching and (university) research. In this context, students are familiarised with the developed software environment, learn the basics of LTA, and work on smaller research projects in practical exercises using the provided digitised material. NumiScience.de will actively integrate the use case database into the service to provide examples of how to handle the data.

Work plan

Duration: 1.5 years

Milestone	Description	Month
1	Concept of data model	1–6
2	Publication specification of data model	6
3	System developed	7–8
4	Data model implemented	9–11
5	Test and release	11–12
6	Concept of LTA interfaces	13–14
7	Publication specification of LTA interfaces	14
8	LTA interfaces implemented	14–16
9	Data ingested in LTA	16–18
10	Overall system released	18

FAIR[1]	F1:RDA-F1-01M / F1:RDA-F1-02M / F2:RDA-F2-01M / F3:RDA-F3-01M / A1: RDA-01-01M / A1:RDA-A1-02M / A1:RDA-A1-02D / A1:RDA-A1-03M / A1:RDA-A1-03D / A1:RDA-A1-04D / A1.2:RDA-A1.2-01D / A2: RDA-A2-01M / I1:RDA-I1-01M / I1:RDA-I1-02M / I2:RDA-I2-01M / I3-RDA-I301M / I3:RDA-I3-02M / I3:RDA-I3-04M / R1:RDA-R1-01M / R1.1: RDA1.1-01M / R1.1:RDA-R1-02M / R1.1:RDA-R1-03M / R1.2:RDA-R1.2-01M / R1.2:RDA-R1.2-02M / R1.3:RDA-R1.3-01M / R1.3:RDAR1.3-02M
TRAILS	related with TRAIL 5.1

[1] Nach Tabelle 1 von Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., … Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Science Journal, 19(1), 41. DOI: http://doi.org/10.5334/dsj-2020-041 cc by 4.0