1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.332

Published: 07 Sept. 2023

FAIRification of Historical Geodata

Automated Metadata Extraction from Archival Maps

Hendrik Herold[1][https://orcid.org/0000-0002-2806-3121], André Hartmann[1][https://orcid.org/0000-0003-0384-8501], Anna Lisa Schwartz[2][https://orcid.org/0000-0003-0391-3879], Michael Hellstern[2], and Markus Schmalzl[2]

[1] Leibniz Institute of Ecological Urban and Regional Development (IOER), Germany

[2] Bavarian State Archives (GDA), Germany

Abstract. A key challenge for the NFDI community is to share and connect datasets across spatial and temporal scales. In particular for Earth System Sciences (ESS) coupling of different RDIs and access to long-term time series data are crucial. In this paper we propose a workflow pipeline for making analogue bound archival geodata FAIR and hence accessible to the different scientific disciplines and reusable in a long-term perspective. We describe an interface for enabling data exchange between an archive (GDA) and a geospatial RDI (IOER Monitor) as well as other RDIs by applying the FAIR data principles.

Keywords: Enabling RDM, Archival Maps, Geospatial Data, FAIR Principles

1. Motivation and Assets of the RDIs

1.1 Archival Geodata (GDA)

The Generaldirektion der Staatlichen Archive Bayerns (henceforth GDA) is the central state authority for all matters of archiving in Bavaria. The Central State Archive of Bavaria and eight regional state archives are subordinated to GDA as technical and administrative head of the Bavarian State Archives. Part of the technical tasks of GDA is the development and maintaining of central digital infrastructures of the Bavarian State Archives, in particular online-services and the Digital Archive. GDA maintains and strengthens a tight network with national and international partners from different research and science communities and the archival community. In close cooperation with research and infrastructure facilities GDA and their associated archives contribute to the development of sustainable information infrastructures in many ways. Among other initiatives and cooperations GDA develops and maintains a digital archive (DIMAG) according to international standards (ISO 14721, PREMIS 3.0) in cooperation with other German state archives, for recording, high-level long-time content preservation and data provenance and providing of born digitals and develops standardised and automated interfaces for the archiving of born digitals. GDA and the associated archives are long-term-archiving, preserving, describing and providing primary analogue and digital research data and corresponding metadata from the collections of state and non-state archives in Bavaria covering 1200 years from 8th century to the present with the aim of open access and use. This includes hundreds of thousands historical maps containing historical geodata from the late 15th to the 20th century.

1.2 LULC-Monitoring (IOER Monitor)

The Monitor of Settlement and Open Space Development (IOER Monitor, https://www.ioer-monitor.de/en/, http://doi.org/10.17616/R3QF5P) is a permanent scientific service within the framework of the research-based Political and Social Consultation Service of the Leibniz Institute for Ecological Urban and Regional Development (IOER). This expert information system is intended to assist scientists, administrators, practitioners from business and industry as well as the general public to answer questions regarding ground cover and land use for the whole of the Federal Republic of Germany. It provides base data for the analysis of land use development, particularly regarding the issue of sustainability. The IOER Monitor is an open research data infrastructure (RDI) in Germany providing domain-specific multi-temporal geospatial datasets, services and visualizations for land use and land cover (LULC)-related development of settlements and open space and closely related topics. Its easy-to-use information system provides multi-scale data offers to form a discussion platform that supports spatial development assessment and evidence-based decision making. It contributes to public land-use change discourses by enhancing information offers that can be adopted by other multi-disciplinary data users - even from non-spatial domains. All data and services are freely available. IOER Monitor is committed to offering continuous services implementing FAIR principles (findable, accessible, interoperable and re-usable) and policy-relevant inputs for transformative spatial development [3]. Figure 1 shows the conceptual framework of the RDI. Historic maps are key to fully describing human influence on LULCC and its impact on the Earth System (cf. [1], [2]), however, they demand high efforts in preparing them for digital analysis.

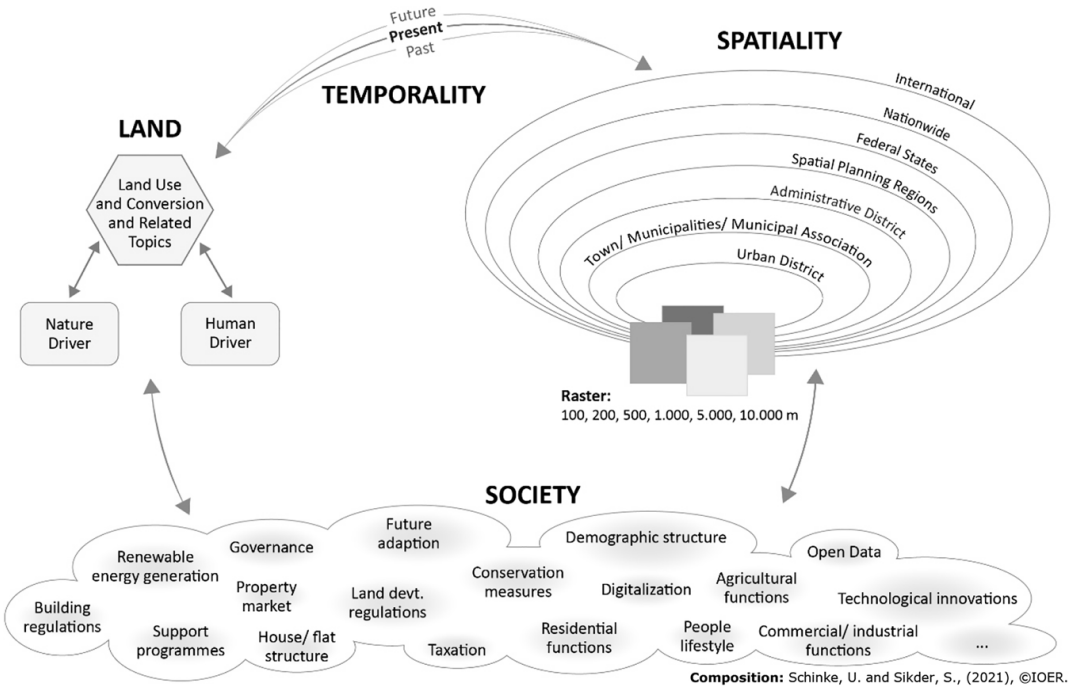

Figure 1. Conceptual framework of the IOER Monitor and the importance of temporality (adopted from [3]).

2. FAIRification of Geodata and Data Life Cycle

2.1 FAIR-Aware Digitization of analogue geospatial archives

Archives and science institutions in Germany and abroad are holding vast quantities of historical maps. For mobilising these analogue geospatial archives the project prepares a best practice workflow starting with Digitization as the first step on the basis of the generalized technical concept of digitization of the Bavarian State Archives. For details see [5] and [6].

2.2 FAIR-Aware Meta-Extraction

We aim at implementing a workflow and data structure, corresponding to FAIR principles (findable, accessible, interoperable and re-usable). Findability-principle is reflected in our general effort to build up a metadata set for archived map scans that can be found, accessed and queried over standard and low hurdle open data pathways. Furthermore, by extracting Map Content, we want to make the information behind, not only the map scan itself, as accessible as possible. This is linked to the more technical part of interoperability, i.e. the use of open standards, formats and software. So by preparing export functionality to a variety of data formats, we allow for a free choice of software on the user side. Finally, the implementation in open source code and software, the definition of an export interface for archiving in addition to the long term archive data provision, enables the reproducibility and reusability of the approach.

Short workflow description (cf. Figure 2):

- QGIS-based Python Toolbox with open source libraries (e.g. OpenCV)
- Map Content Extraction for Meta Data enrichment
- Export of map content and metadata to non-proprietary data formats (e.g. XML, GML, NAS, INSPIRE-conform geodata)

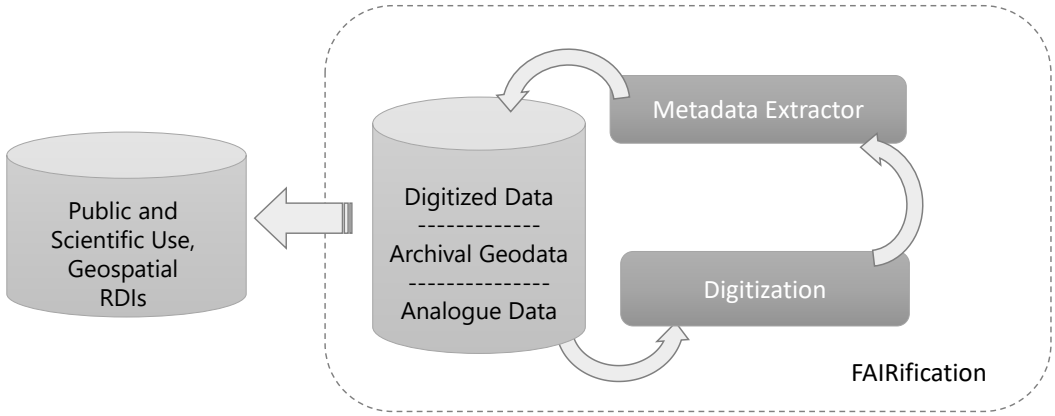

Figure 2. Simplified concept of the proposed pipeline.

2.3 FAIR-Aware Data Life Cycle

The project aims at the FAIRification of historical geodata covering the whole data life cycle. By this means metadata is extracted in interoperable and standardized data and metadata formats and enriched with metadata suitable for current interoperability as well as long term archiving.

For archiving and ensuring the reusability of data the project uses the Generalized XML-Client of GDA addressing the current gap in securing the data life cycle. This allows for automatized structuring, ingesting and description of the data and secures data provenance by documentation of all individual operations as well as the implementation by other repositories. It requires a structured Submission Information Package and a well-documented XSD-file. Within the project archival formats and a list of required data and metadata for documentation for the long-term archiving and reuse of the data are defined as a best practice. Based on the generalized XML-client an interface is aspired for the automated output of all relevant data and metadata.

Data availability statement

The results of this project are metadata sets, which also have a geodata component. The machine-readable textual part of the metadata is relevant for archive and catalogue systems, and will be available likewise. The geodata part of the metadata will be additionally available in a future research data infrastructure. The code and data produced will be made freely available (GitHub, https://github.com/ioer-dresden, etc.).

Underlying and related material

The above described RDI IOER Monitor is available here: http://doi.org/10.17616/R3QF5P

Author contributions

Hendrik Herold contributed in Funding Acquisition, Conceptualization, Investigation, Data curation, Software, Writing – original draft, Writing – review & editing. André Hartmann contributed in Software, Formal Analysis, Data Curation, Writing – review & editing. Anna Lisa Schwartz contributed in Conceptualization, Data Validation, Writing – review &editing. Michael Hellstern contributed in Conceptualization, Data Validation. Markus Schmalzl contributed in Funding Acquisition, Conceptualization, Investigation, Data curation, Software, Writing – original draft, Writing – review & editing.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was partially funded by NFDI4Biodiversity (Flexfund-Project P-06/2022).

Acknowledgement

The authors would like to thank NFDI4Biodiversity and DFG for funding parts of this work.

References

1. H. Herold, “Big Historical Geodata for Urban and Environmental Research”, In: Handbook of Big Geospatial Data. Springer, (2021), pp. 475-486, doi: https://doi.org/10.1007/978-3-030-55462-0_18
2. H. Herold, “Geoinformation from the past – computational retrieval and retrospective monitoring of historical land use,” Wiesbaden, Springer, 2017, 192 p. http://dx.doi.org/10.1007/978-3-658-20570-6
3. G. Meinel, S. Sikder, T. Krüger, „IOER Monitor: A Spatio-Temporal Research Data Infrastructure on Settlement and Open Space Development in Germany”, In: Jahrbücher für Nationalökonomie und Statistik 242, 2022, 1, pp.159-170, doi: https://doi.org/10.1515/jbnst-2021-0009
4. Holzapfl, Julian, et al. Quick wins und dicke Bretter. Übernahme und Archivierung von Fachverfahren. Archiv. Theorie & Praxis, vol. 76, no. 1, 2023, pp. 15-24.
5. Puchta, Michael. Automatisierung und Standardisierung. Ein Praxisbericht aus den Staatlichen Archiven Bayerns. Archiv. Theorie & Praxis, vol. 74, no. 3, 2021, pp. 180-186.

6. Puchta, Michael, et al. Fachkonzept für das Digitale Archiv der Staatlichen Archive Bayerns. Staatliche Archive Bayerns, München 2023, doi: https://doi.org/10.5281/zenodo.7743888.