NFDI4Earth

Interoperability and reusability of geoscientific lab data

Matthias Halisch[1], Andreas Weller[2], Francois Holtz[3], and Monika Sester[4]

[1]Leibniz Institute for Applied Geophysics, Stilleweg 2, D-30655 Hannover
[2]Institute of Geophysics, TU Clausthal, Arnold-Sommerfeld-Str. 1, D-38678 Clausthal-Zellerfeld
[3]Institute of Mineralogy, Leibniz-Universität Hannover, Callinstr. 3-9, D-30167 Hannover
[4]Institute of Cartographics and Geoinformatics, Leibniz-Universität Hannover, Appelstr. 9a, D-30167 Hannover

30.07.2020

Abstract

Geoscientific research has become a highly complex and interdisciplinary task that often needs several datasets to answer complex questions, and that produces huge amounts of manifold data, associated data types and related documentations accordingly. Although a broad range of possibilities exist to make these data findable and accessible (e.g., by assigning digital object identifier and by storing data in repositories), interoperability and reusability are mostly not guaranteed or even not possible, due to non-existing interdisciplinary standards and highly variable scales of research (i.e., from laboratory to fieldwork). Accordingly, the main objective of this pilot-project is to develop an urgently needed interoperability and reusability framework, including inherent and meaningful standards, by exemplarily utilizing and assessing a broad range of available interdisciplinary ***geoscientific laboratory data*** *and data types* ***across different scales*** *(i.e., petrophysical, geological, mineralogical and image-based data), with regard to hydraulic transport in porous rocks. Based on these different datasets and data types, this pilot will develop methods and assign standardized metadata, in order to ensure a most complete data integration for many independent research questions exemplarily. A web-based platform will be set up, to demonstrate, distribute and provide the developed standards and research data as intended for the NFDI in general, and for the NFDI4Earth in particular. This project primarily aims at scientists, data curators, university teachers, and decision-makers; results are also relevant for infrastructure providers and system integrators for the development of new or the improvement of existing archives and repositories. Based on datasets created, compiled and exchanged within the FZ:GEO research network (Leibniz Forschungszentrum GEO at Leibniz Universität Hannover) and associated project partners, this pilot-project will create the basis for interoperability and reusability of geophysical and geochemical datasets. For the future, such web-based interface could be easily extended towards arbitrary fields of geoscience, including arbitrary scales of research and will become a powerful tool for providing information and data for new research concepts and fields.*

I. Introduction

Geoscientific research very often needs and relies on extensive interdisciplinary work, in order to provide more and more complex answers for an increasingly wider range of stakeholders. This pilot addresses an important scientific challenge, namely the prediction of hydraulic properties of consolidated and unconsolidated rocks, and shows, which disciplines and which datasets are required for solving this task. In order to analyze hydraulic and mineralogical properties fast and reliably, the joint integration and analysis of three different types of data are relevant, including image-based and classical laboratory data. This NFDI4Earth pilot will develop processes and methods for a joint data integration and enrichment, in order to make these exemplary datasets NFDI-ready, as well as to fulfill the demands of the recently launched “Geologiedatengesetz” (GeolDG). It will focus upon the automatic determination of metadata, which will describe all datasets sufficiently, in order to perform a search query, yielding comprehensive information about the data with respect to specific scientific questions. This task also includes mechanisms for a harmonization of keywords and notions used by the different geoscientific disciplines (e.g., geophysics, geology, hydrology, mineralogy, etc.).

Each of the three exemplary datasets that have been selected for the pilot features highly individual characteristics and challenges:

- The *petrophysical datasets* typically consist of results from multiple methods and very often feature “redundantly” derived parameters (e.g., porosity, which can be determined amongst others by triple weighing, nuclear magnetic resonance spectroscopy, pycnometry, mercury intrusion, gas adsorption). All methods have different technical drawbacks, limitations, use different sample sizes (mm to cm) and resolutions (nm to mm), leading to slightly different parameter variations from the very same investigated specimen. Hence, providing complete documentation and metadata is essential for the “understanding” and hence for the reusability of these datasets. It is common use, to compile sample and / or site location data within a single (Excel-) worksheet, which leads to a huge amount of files to be assessed and curated.
- The *3-D image-based datasets* typically consist of very large data (up to 100 GB per sample!), available in very different data formats (e.g., 3-D raw data, 2-D image stacks, etc.), featuring different image resolutions (nm to mm) and investigated sample volumes (mm^3 to cm^3). Since these large datasets are very challenging to store, it is common practice to delete the image data after processing and publication and to archive the physical sample instead. With detailed documentation of the boundary conditions of the scan, the original 3-D dataset can be reproduced if needed. Nevertheless, indexing and assigning metadata is very often missing.
- The *mineralogical datasets* can provide the chemical composition of solid phases in Earth Sciences (mainly minerals). Samples are usually characterized in 2-D with a variety of routine methods (e.g., Back Scattered Electron images) at different scales and resolutions. However, the storage and further processing of the analytical image data remains challenging, especially the integration of many high resolution images for the description of large samples (so called “stitching”). Besides, the extraction of quantitative information of chemical compositions and data is often impossible, because analytical protocols, documentations and workflows are not specified. As a result, a huge amount of data on element distribution in minerals remains unexploited and hence is not available for further processing tasks of other scientists.

The pilot concentrates on setting a framework and standards for true interoperability and reusability of large and joint datasets, which highly promises that a tangible result is reached. The problems that have to be addressed here are general enough to allow for a subsequent extension to other applications, analytical methods and research domains.

II. Pilot Description

Within this pilot project, an exemplary web-based interface will be set-up conceptually, developed, tested and assessed for classical laboratory data (discrete numerical value, mostly ASCI-file format), as well as from 2-D and 3-D image-based data (2-D and 3-D images, typically 2-D images or 3-D raw files). For these task, machine learning based methods and algorithms will be utilized (e.g., convolutional neural networks, matching, sequence mining, etc.). For the description of the data, well known standards of the open geospatial consortium (OGC) in general and of the catalog service for the web (CSW) in particular, will be applied. The developed software / interface will rely on publicly available frameworks and will be released and made accessible as open source software using creative commons license (CCL). For ensuring the interoperability of the pilot, an integrative approach will be used, i.e. based upon the available interdisciplinary data, detailed descriptions will be derived that will be used to develop feasible and common standards. In order to reach these goals, three exemplary datasets, available from other recent projects, will be utilized and upgraded, in order to be NFDI-ready. Based upon the techniques as stated before, metadata will be extracted automatically. A concept for the determination of meaningful “data units”, with special regards to different scales of investigation (e.g., a section of a borehole core, a handpiece, a core plug, small fragments, etc.) for an unambiguously data assignment will be developed. Furthermore, a harmonization mechanism for handling subject-specific descriptions by different geoscientific communities will be implemented: firstly, by expert interviews, secondly, by automatic tools, which allow to infer the relationship between concepts based upon the investigation of the data. With this pilot, a conceptual framework and a practical toolbox will be developed that will bridge subject-specific data hurdles and hence deliver a sustainable and true interoperability and reusability potential for geoscientific data.

III. Relevance for the NFDI4Earth

This pilot is primarily addressing the following stakeholders: scientists, data curators, university teachers and researchers, since these groups directly represent not only the “producers” of data, but furthermore the main group for interdisciplinary data (re-)usage and the creation of projects and hence of new knowledge thereupon. This new knowledge might target at decision-makers and the public authority for the future. Secondly, this pilot addresses other infrastructure providers and system integrators, not only to implement the developed standards, but also to utilize the developed interface for an enhanced and prospective integration of highly complex research data. From our starting point, this project will first utilize joint data from the fields of petrophysics, geology / mineralogy, as well as from 2-D and 3-D image based material analysis of consolidated and un-consolidated natural material mostly derived at the laboratory scale. By concept, the interface will be easily extendable to any other geoscientific research field, scale and location within the Earth System Sciences. As pointed out earlier, we particularly target the elements “interoperability” and “reusability” of the “FAIR-principles”, since these elements, by the best of our practice and knowledge, urgently need standards and guidelines to make research data “truly FAIR”. For this, we need to cover the entire span of the “research data life cycle” within this pilot, in order to ensure sustainable cross-discipline reusage of research data especially. Alltogether, this pilot will deliver valuable results for advancing interoperability within the NFDI4Earth. The interface and platform to be developed will help to increase the reusability of available data across geoscientific disciplines significantly. By linking this interface to already existing data infrastructures, a significantly improved joint data integration will be achieved.

IV. Deliverables

IV.a Technical Operability

The pilot will deliver a developed and implemented pipeline for processing the mentioned datasets into a NFDI-ready form. All created software will be open source. The extensibility and transferability of results will be exemplified with at least three other comparable data sources. Numerical, as well as web coding procedures will be performed by preferably using well-established and freely available software packages (e.g., Python, Keras, Bootstrap). Developed codes and algorithms will be fully documented, and provided on the project website. The prototype of the web-based data interface will be tested within the FZ:GEO research network and will be fully accessible after the end of this pilot. The LIAG will cover all following costs for keeping and maintaining this interface. The progress and outcome of the pilot will be communicated frequently and transparently within the related scientific community. The project itself will be used for the education of young researchers with special regards to good scientific practice.

IV.b Roadmap Title

According to the basic idea of this pilot, the roadmap for follow-up work will possibly be entitled: ***“Making geoscientific laboratory data truly “FAIR” – bridging disciplines and setting standards for sustainable research data usage”.***

V. Work Plan & Requested Funding

Project tasks & milestones	year 1							
	Q1.1	Q1.2	Q2.1	Q2.2	Q3.1	Q3.2	Q4.1	Q4.2
Pilot-Project @LIAG								
1.1: Joint data assessment and curation	▨	▨						
1.2: Development and assignment of interdisciplinary metadata		▨	▨					
1.3: Development of joint data units according to FAIR-principles		▨	▨	▨				
1.4: Development of an exemplary web-based interface				▨	▨	▨	▨	▨
1.5: Project management, scientific transfer & outreach	░	░	░	░	░	░	░	░
M.1: Concept for joint data interoperability and reusability finalized				■				
M.2: Web-interface complete and joint data implemented							■	■

Legend: 1.1 – 1.5 = project tasks; M.1 & M.2 = milestones.

According to the boundary conditions of the NFDI4Earth, we apply for a post-doc data scientist, based upon the funding list of the Deutsche Forschungsgemeinschaft (DFG): 100 %, TV-L, E14, 74100 € p.a.