BPMN4EARTH: Metadata Enriched & Automated Workflows

Gunnar Pruß[1] and Oliver Rach[1]

[1]Helmholtz Centre Potsdam – German Research Centre for Geosciences GFZ, Section 4.6 Geomorphology, Telegrafenberg, 14473 Potsdam, ROR: 04z8jg394

March 31, 2023

Abstract

The FAIR principles - Findability, Accessibility, Interoperability and Reusability - form the basis of modern research data management and are crucial for laboratory environments with their wide array of administrative, descriptive, structural, and technical metadata. Currently though, we are lacking a generic solution to capture these metadata in the Earth System Science (ESS) domain where complex and problem driven workflows are inherently difficult to standardize. In BPMN4Earth, we want to use the Business Process Model and Notation (BPMN), a standardized tool that is already established in many authorities and companies, to make workflows in our laboratories comprehensible and reproducible. BPMN enables a flexible and comprehensive workflow description that can be visualized for eased communication and publication while simultaneously being machine-readable and hence suitable for process automation. With BPMN4Earth, we intend to develop an open source application to i) create, search, and combine metadata enriched workflow descriptions for laboratory environments based on BPMN, ii) obtain persistent identifiers (PID) to reference workflow descriptions in publications, and iii) automate data analysis workflows like calibration, evaluation, aggregation, and visuali-zation. The pilot application initially targets laboratories at the German Research Centre for Geosciences (GFZ). In general BPMN4Earth be used to describe any complex data generation process and enrich it with metadata, but specifically, it addresses the needs of laboratories that lack well defined standard operating procedures (SOP), or for which the scope of existing electronic lab notebooks (ELN) is too limited in an ESS context.

I. Introduction

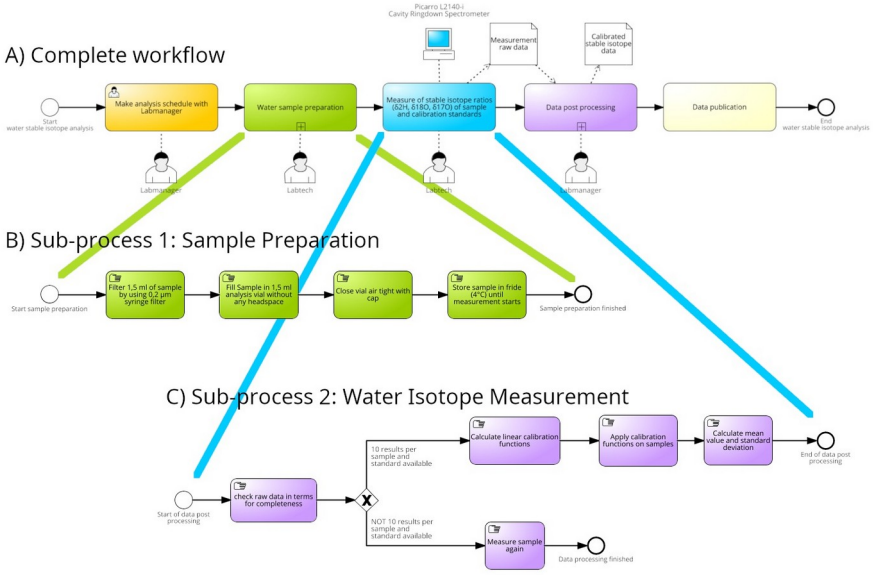
In the Earth System Science (ESS) domain, we are confronted with an ever growing amount of (meta-) data from samples, observations/measurements, and analyses. The demand for high spatial and temporal resolution inevitably also leads to an increasing number of samples that need to be processed, analyzed, and evaluated in our laboratories. The explanatory power and reusability of data generated in a laboratory is highly dependent on the description of laboratory workflows though, namely the documentation of i) sample preparation methods (which steps have been applied to a sample), ii) instrument settings (e.g. temperature profile,

split-volume, injection-volume), and iii) subsequent data processing (e.g. calibration, evaluation, aggregation, and visualization). In practice, workflow descriptions are often implicitly stored as a combination of analogue lab notebooks, text and spreadsheet documents, wiki applications, and increasingly electronic lab notebooks (ELN). Therefore the available data and metadata on workflows varies widely with respect to structure, format and level of detail. Wikis, for instance, offer a lot of flexibility, but can neither be referenced in publications nor be automatically processed. ELNs on the other hand often have a limited scope, or miss crucial features or APIs that are required to integrate them into the complex, interdisciplinary settings of ESS research with its wide range of different sample types (e.g. rock, sediment, soil, water, and plant material) and analytical methods (e.g. grain size measurements, stable isotope and biomarker analyses, cosmogenic nuclide dating, and water chemistry). There are some very recent suggestions from the German Marine Research Alliance to publish workflows for routine operations as standard operating procedures (SOP) (e.g. Thomisch et al., 2023 and Getzlaff, 2023), but these are currently only available as textual descriptions in PDF format. These shortcomings of current solutions have a negative impact on comprehensibility, data availability, and reusability, and they also affect laboratory efficiency and on-boarding procedures for new staff and visiting scientists. With BPMN4Earth we want to develop a workflow description tool that is highly flexible and both visualizable and machine-readable. We want users to (literally) see the processing steps that have been performed in the laboratory, and understand which data their results are based on, as well as the assumptions that were incorporated. When the relevant metadata are contained in workflow descriptions, it also becomes possible to use descriptions as dependency graphs and conduct sensitivity analyses to create awareness at which processing steps special diligence is required. Furthermore, we would like users to be able to split complex workflows into manageable components (lab methods, data acquisition, data storage, data analysis, publication, etc.) with well defined inputs and outputs which can be individually developed, optimized, configured, combined/reused, and (if applicable) automated. The identified (sub-) components could then be assigned to the most qualified personnel available (student helpers, lab technicians, lab managers/heads, data stewards, (data) scientists, etc).

II. Pilot description

With BPMN4Earth we suggest to use Business Process Model and Notation (BPMN) for workflow descriptions in ESS laboratories. BPMN is an internationally established, well documented and widely used standard for process modeling that is based on the Extensible Markup Language (XML). To users BPMN workflow descriptions are usually presented in a graphical format (see figure) that is easy to comprehend and communicate, because it is based on a relatively small set of graphical elements (events, activities, gateways, and connections). Each object in a workflow, e.g. a task, can be associated with any number of metadata

attributes which makes BPMN highly flexible and suitable to model even complex laboratory processes. Moreover, once a workflow has become too complex, or a recurring pattern has been detected, it can be broken down into comprehensible and reusable sub-processes.



In addition, the workflows are machine-readable and events like a finished batch analysis in a laboratory can for instance trigger actions in a data processing pipeline such as quality control or calibration. In our pilot we will create BPMN workflows for the qualitative and quantitative analyses of biomarkers, compound specific stable isotope analysis, and water isotope measurements conducted in the Section 4.6 (Geomorphology) at GFZ. We'll then identify recurring patterns to mark them as SOPs and register them with the Library and Information Services (LIS) at GFZ to make them accessible via persistent identifiers (PID). Within BPMN4Earth, we'll duplicate as little functionality and data as possible and make use of existing APIs offered by LIS, ORCID, ROR, already established sensor management systems, (e.g. SMS at the GFZ), and laboratory infrastructure portals (e.g. LI2 by the Geo.X network). The software components designated for BPMN4Earth are open source, operating system independent, and mostly based on the Python ecosystem. We are looking into Snakemake and SpiffWorkflow for automation and workflow management, as well as topic specific data analysis and machine learning libraries. For the browser-based application we will use a simple Django front end with a PostgreSQL database as back end. In both cases the open source JavaScript library bpmn-js will be used for BPMN browser integration. All source code, as well as the our BPMN workflows (xml), will be put under version control with git. We'll share BPMN4Earth early on via GitLab and encourage all interested parties to give feedback, raise concerns and request additional features.

III. Relevance for the NFDI4Earth

Detailed information about laboratory processes and analysis methods is an essential component of the metadata used to generate research data in the laboratory. The possibility of mapping this information in a comprehensible way by means of an already existing international standard and storing it in a machine-readable (and thus executable) form enables sustainable and long-term use of the research data in terms of the FAIR principles interoperability and reusability. In particular, our pilot project enables data providers, curators and users to store research data with extensive metadata and to better evaluate the information provided by third-party data for their own research purposes. After the implementation and test phase of the pilot in the laboratories of the Geomorphology section, test phases in other geochemical laboratories within the GFZ are planned. In addition, presentations and at least one scientific publication are planned. Within the framework of the "Analytics GFZ" laboratory community, lectures and seminars on laboratory-related topics will be held on a regular basis. Presentations are planned in this seminar series as well as at conferences such as the European Geophysical Union (EGU) where the appropriate audience (scientists, data curators) is present. The manifold use of BPMN (among others in the context of SAP) in many internationally active companies for the visualization and automation of production and decision processes shows the universal possibilities of this tool. After a successful implementation of the pilot in the geochemical domain, we consider an extension to other scientific domains as absolutely possible and even recommendable.

IV. Deliverables

- WP1: BPMN workflows: 1) Become acquainted with BPMN, study sample applications and best practices, 2) Identify a set of exemplary preparative, analytical and post-processing methods, 3) Translate methods into BPMN workflows, 4) Convert recurring patterns into sub-processes with well defined inputs and outputs.
- WP2: PIDs for workflows - Create a pipeline to obtain PIDs for (sub-) processes.
- WP3: Automation: 1) Identify and select processes that are suitable for automation 2) Automate a representative subset of the selected processes, 3) Create example pipelines that orchestrates manual and automated processes.
- WP4: Metadata export: 1) Export metadata for individual SOPs, 2) Export metadata associated with chained processes (data pipelines).
- WP5: User interface: 1) Create a basic browser-based application for users, 2) Create documentation.
- WP6: Publication and outreach: 1) Multiple discussions and presentations at Analytics GFZ, 2) Presentation at EGU, 3) Scientific publication, 3) Roadmap document: *BPMN4Earth - Metadata Enriched & Automated Workflows*

V. Work Plan & Requested funding

Month	1	2	3	4	5	6	7	8	9	10	11	12
WP1: BPMN workflows												
WP2: PIDs for workflows												
WP3: Automation												
WP4: Metadata export												
WP5: User interface												
WP6: Publication and outreach												

We request funding for a one-year full time equivalent for Gunnar Pruß for a scientific position (TVL/öD E13, level 3) which is (based on the Personalmittelsatz of the DFG for 2023) 80,100.00 Euro. No material resources are requested.

VI. References

- Getzlaff, Klaus. (2023). SOP Model Data Submission to THREDDS at GEOMAR (1.3). Zenodo. https://doi.org/10.5281/zenodo.7674489
- Thomisch, Karolin, Spiesecke, Stefanie, & Boebel, Olaf. (2023). Standard operating procedures: Featuring Passive Acoustic Data by The Open Portal to Underwater Soundscapes (OPUS), Part I: Data preparation and standardization (1.0.1). Zenodo. https://doi.org/10.5281/zenodo.7680029