4D-WORKS: DEVELOPMENT OF STANDARDIZED METADATA FOR 3D TIME SERIES WORKFLOWS

Katharina Anders[1]

[1]Professorship of Remote Sensing Applications, TUM School of Engineering and Design, Technical University of Munich, Germany

June 22nd, 2024

Abstract: Topographic 3D time series acquired through laser scanning or photogrammetry become increasingly available in many domains of Earth System Sciences. While standards and best practices have been established for 3D data, they are mostly lacking for multitemporal or time series data, where processing is currently mostly time-agnostic and (meta)data handled in a user-specific design. The objective of this pilot project is to develop a curation workflow for 3D time series data from different sources (platforms and sensors). Particular focus is on the development of standardized time-dependent metadata that can be automatically generated from different data acquisition strategies. The result will be a comprehensive guide and tutorials about developed best practices as a foundation for metadata standards, provided with an open-access repository of template material and scripts to integrate in workflows of the user community. Users are researchers and practitioners from any domain in Earth System Sciences that use multitemporal topographic data, i.a. from geomorphology, hydrology, ecology, and public and private organizations working in environmental monitoring contexts. The pilot project will strongly facilitate access and (re)usability of 3D time series for prospective users, where the developed metadata-rich workflows are fundamental for efficient and flexible use in domain-specific tasks. This advances RDM workflows to fully consider the special properties of 3D time series (4D data) and making them accessible to the wider user community. This contributes to NFDI4Earth efforts especially regarding long-term perspectives of data use and efficiency of analysis workflows.

1 Introduction

1.1 Which of the track(s) are you applying to?

Data Curation Pilot – Enhancing data quality, interoperability, and reusability within ESS through the development of automated curation workflows, tools, and assistance mechanisms.

1.2 What is the data-challenge you face and what is the current state?

4D topographic observations, i.e. dense **3D time series**, are acquired using laser scanning or photogrammetry to assess changes in a scene throughout different points in time. Compared to traditional multitemporal datasets that are generated with repeat point cloud acquisitions at seasonal to (multi-)annual scales, the time series consist of a high number of epochs (typically hundreds to thousands; e.g., Vos et al. 2021, Anders et al. 2023). Regarding typical processing workflows, this poses a **great challenge to consistently handle these data and associated metainformation**. There is no standard or established practice regarding time-dependent metadata, starting from the timestamp of acquisitions. **Essential processing information and products**, such as transformation matrices required for time-dependent alignment of epochs, is not handled in a consistent manner and usually provided as supplementary files in **formats tailored to the user-specific tools** that were used for the processing. Reusing the data in different workflows and domains is tedious, as they often require customized scripts to handle the large volumes of 4D (3D+time) data. Especially with emerging efforts of research groups to publish such valuable datasets, the **development of uniform and flexible metadata and processing practices** is required to lay an early foundation of high interoperability and reusability of 4D data.

1.3 What is your vision for your community and their RDM workflows if your challenge would be solved?

If this challenge were solved, RDM workflows by the community would become **much more efficient**, as data would be highly interoperable and could be easily integrated by different tools and with clearly and openly defined metadata interfaces. Furthermore, solving the lack of metadata practices for 4D data could lead to **more datasets being published**, as the effort decreases for publishers in preparing their data in an accessible way (without any structured information how this can be done), on the one hand, and on the other hand the potential increases that **published data will be used by others**.

2 Pilot description

2.1 What is the proposed solution to your thematic track?

To solve the described challenge, we propose to **develop a metadata format and automatic curation workflow** for point cloud epochs with special focus on time-dependent information. To ensure that the developed practice will be employed by the community, an automatic workflow is required to generate the metadata for **representative use cases** of laser scanning and photogrammetry time series acquisition. Both the respective acquisition characteristics (LiDAR and photogrammetry formats) as well as properties of commonly deployed manufacturers and software will be considered for automatic curation of metadata in the harmonized format.

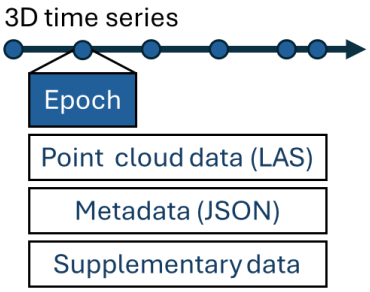

Figure 1: Schematic of 3D time series with each epoch consisting of point cloud data, a metadata file and (optionally) further supplementary data. The standardized metadata is automatically generated for a range of acquisition strategies and can be integrated into workflows by reading into different tools. Temporal sampling of epochs can be irregular, and data acquired by different sensors.

2.2 What is the technological backbone you rely upon?

The curation workflow will be implemented in an **open Python library**, which is openly developed and documented on **GitHub**. Documentation and tutorials will be published with **automatic website generation** via GitHub pages. The repository will be released and versioned via Zenodo (with DOI). The metadata **format and schema are foreseen** using JSON or YAML files, both simple data formats with readable text, which support different data types used by dynamic programming languages. The final choice of format will be part of the research in the beginning of the pilot. Advantage of YAML is that it supports date formats and is widely used for configuration of automated services, whereas JSON may provide a higher compatibility with existent tools.

2.3 What are the standards and interoperability approaches used in the pilot's context?

The developed metadata schema and format will be **based on existent standards**, such as LAS for laser scanning point clouds (ASPRS 2019) or the quite recent Open Photogrammetry Format (OPF; Krahenbühl et al. 2023). So far, existent standards mostly define **acquisitional metadata** that does not describe temporal information (as it is not treated as epoch of a multitemporal dataset/time series). The schema will consider standards used in satellite remote sensing, from where useful approaches may be adopted (e.g., Open Data Cube). Metadata describing the processing of epochs (e.g., time-dependent transformation matrices) will build on standards by **established open source software**, such as PDAL. Interoperability is ensured by adhering to existent definitions and strongly linking to established procedures while warranting the **use of exchangeable data formats**. Where required, the developed curation workflow will contain interfaces to important software tools (e.g., PDAL, CloudCompare, py4dgeo) to ensure compatibility with existent workflows. Data interoperability will be fully maintained, as no existent metadata standards will be replaced, but entirely extended by the lacking time series and time-dependent processing component.

2.4 How will your pilot produce impactful showcases of usability within the German Earth System Science community?

An impactful **application case** of photogrammetric and laser scanning data acquisition will be produced at a new monitoring site of river dynamics of regional environmental importance (see 4.2). Thereby, a valuable 4D dataset will be acquired and curated for publication according to the pilot's solution. Furthermore, **existent data publications** with comprehensive but user-specific metadata descriptions (Vos et al. 2021, Anders et al. 2023) will be automatically translated to the new metadata definition, presenting the ease and value of the workflow for both future data acquisitions and legacy data. The applicability of the pilot will be showcased with a full analysis workflow of 4D point clouds (using the application case and existent data) with the scientific open-source software py4dgeo (py4dgeo Development Team 2024), a Python library to bundle methods for change analysis in 4D point clouds.

3 Relevance for the NFDI4Earth

3.1 What are expected users and stakeholders and how do they benefit from your solution?

Expected users are **scientists and data curators** who benefit from the standardized metadata formats both when

acquiring and processing own data much more efficiently as well as using published data, which so far can often be integrated in own workflows only by preparing tailored scripts in a first step. The latter is also highly relevant for **university teachers**, where diversification of use cases, e.g., in practical courses is strongly facilitated by using harmonized (meta)data. **Infrastructure providers**, such as data repositories, can use the standard and workflow as instrument to perform technical checks on submitted data, and importantly, to provide guidelines/requirements for publishers of 3D time series data. Ultimately, the metadata standard and curation workflow benefit **all end users** (including public authorities and decision makers) who rely on transparent and reproducible workflows to further process and interpret information derived from 4D data in Earth observation and environmental monitoring.

3.2 What measures are planned to support the uptake of your solution in your community, and to engage with prospective users?

The solution will be **discussed with the community** in the early stages and during development for feedback at national and international conferences (e.g., DGPF, ISPRS, EGU). A **workshop** will be conducted upon completion of the project to provide a low-threshold entry for users. This will be based on **online training material** (tutorials) in the GitHub repository with permanent accessibility. The **issue tracker system** of GitHub will be used for active engagement in form of questions, feedback, or new suggestions.

3.3 What is the potential for other ESS subdomains?

The challenge and proposed solution are relevant for a **variety of ESS subdomains** which use topographic time series for observing and/or modelling surface dynamics and related processes. Prominent examples are geomorphology, cryospherology, hydrology, and ecology (for vegetation dynamics). The developed metadata approach may provide a **basis for other time series observations** with similar requirements, and thereby support RDM development beyond the scope of this pilot's data and application domains.

3.4 How does your pilot enact FAIR principles?

The pilot fundamentally enhances the **accessibility** of 3D time series data in terms of harmonized and interpretable metadata. **Interoperability** is increased for different workflows and tools, which are heterogeneous but relatively limited to a certain range of established software. Thereby, **reuse** of these powerful but challenging datasets is encouraged through reducing efforts or even removing obstacles (such as modifications being non-reversible due to lack of processing information). Given the planned structure and provision of the solution, the FAIR principles are fully enacted for the pilot itself and constitute the **primary objective** in order to ensure uptake in the community.

3.5 What challenges in managing earth system research data does your pilot address?

The pilot addresses the challenges of (re)using and processing **large data volumes** with heterogeneous and **domain-specific workflows**, and the challenge of preparing data for archiving and publication with respect to the FAIR principles with a lack of metadata standards for 3D time series.

3.6 Are there particular contributions that help the NFDI4Earth to engage with e.g., analysis methods, education material, advancing interoperability, or new platforms, ...?

The tutorials and application case may help to engage in data curation and **analysis methods for 4D point clouds** or photogrammetric and LiDAR data in general. The planned GitHub repository can serve as **good practice example** for the development of other curation workflows with comprehensive documentation and living platform for engaging users.

4 Deliverables

4.1 Technical operability of the pilot

The pilot will operate in an open **GitHub repository** where all material is developed and shared consisting of a **Python library** and extensive metadata templates. Upon completion of the one-year project, the repository will be archived as first version to **Zenodo with DOI**.

4.2 Material or actions for dissemination of knowledge/data

The pilot project includes an **application case** of multi-source data acquisition (UAV photogrammetry, terrestrial laser scanning, UAV laser scanning, (experimental) permanent laser scanning) of sedimentation dynamics conducted on the Isar river at a study site of regional environmental relevance. Acquisitions are planned in August

and October 2024; collaboration with local stakeholders are established, and all permits granted. The full dataset will be **openly published** (planned on the PANGAEA data repository) based on the curation workflow of this pilot project. The application case will be prepared for publication as **peer-reviewed article** in an open-source Earth System Science Data journal (e.g., Copernicus ESSD). A comprehensive **documentation** of the curation workflow, including **tutorials** and **demos** (application case) will be created within the GitHub repository with automatic website generation (GitHub pages). This concept allows to extend the project in the future (e.g., further tutorials, additional tools, new demos) and to improve and extend the workflow and resources with automatic updating of all material. A **workshop** will be offered embedded in a relevant event of the user community, e.g., the EGU General Assembly or an ISPRS workshop/congress upcoming by the end of 2025/2026 (in-kind contribution of travel costs).

4.3 Roadmap document for the community, including project evaluation and future directions

A **roadmap document** titled "Towards metadata standards and automatic curation workflows for 3D time series data considering interoperability and reusability in open-source tools and standard software" will be prepared and published with the project repository (incl. DOI via Zenodo archive). Integration in the NFDI4Earth Living Handbook is aspired.

5 Work plan & requested funding

5.1 Provide a milestone plan (or Gantt chart) for the planned one-year implementation phase.

Work packages (WP)		1	2	3	4	5	6	7	8	9	10	11	12
WP1	Research of metadata practices in existent (published) datasets, software and literature	■											
WP2	Conceptual design of metadata requirements and implementation schemes		■	■									
WP3	Development of automatic workflows for metadata generation and handling			■	■	■							
WP4	Implementation of curation workflow (development and testing in open repository)					■	■	■	■				
WP5	Implementation of application case from data acquisition to analysis results								■	■	■		
WP6	Dissemination of knowledge and data (publications, roadmap, workshop preparation)										■	■	■
Milestones (M)													
M1:	Metadata concept and automatic curation workflow developed					X							
M2:	Curation workflow implemented, pre-release of open-source repository complete								X				
M3:	Dataset (application case) published on PANGAEA											X	
M4:	Roadmap published via Zenodo, NFDI4Earth integration organized												X
M5	Scientific article submitted to ESSD journal												X
M6:	Workshop for user community organized for end of 2025/2026												X

5.2 Write how much money and what position you are applying for.

We apply for a **total amount of 83,874 €** which covers funding of a one-year full-time position (12 PM) for a doctoral researcher (TV-L E13) in the amount of 79,800 € (based on DFG personnel rates for 2024) and 4.074 € travel costs for participating in the EGU General Assembly 2025 and in the NFDI4Earth plenary for two persons (PI and project researcher), respectively. All infrastructure is provided by the university and seed funding of the professorship. Travel costs for data acquisition and conducting a workshop are contributed in-kind.

References

Anders, K., Eberlein, S., Höfle, B. (2023): Hourly Terrestrial Laser Scanning Point Clouds of Snow Cover in the Area of the Schneeferner, Zugspitze, Germany. PANGAEA. https://doi.org/10.1594/PANGAEA.941550.

ASPRS (2019): LAS Specification 1.4 - R15. URL: https://github.com/ASPRSorg/LAS (2024-06-22).

Grégoire Krähenbühl, Klaus Schneider-Zapp, Bastien Dalla Piazza, Juan Hernando, Juan Palacios, Massimiliano Bellomo, Mohamed-Ghaith Kaabi, Christoph Strecha (2023): The Open Photogrammetry Format Specification. Pix4D, 2023. URL: https://pix4d.github.io/opf-spec/ (2024-06-22).

py4dgeo Development Core Team. (2023). py4dgeo: library for change analysis in 4D point clouds. (Version 0.6.0) [Computer software]. https://github.com/3dgeo-heidelberg/py4dgeo.

Vos, S., Anders, K., Kuschnerus, M., Lindenbergh, R., Höfle, B., Aarninkhof, S., de Vries, S (2021): A six month high resolution 4D geospatial stationiary laser scan dataset of the Kijkduin beach dune system, The Netherlands. PANGAEA. https://doi.org/10.1594/PANGAEA.934058.