AGILE: GIScience Series, 2, 31, 2021. https://doi.org/10.5194/agile-giss-2-31-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

Recommendations for Future Data Management Plans in Earth System Sciences

Christin Henzen[a] (corresponding author), Stefano Della Chiesa[a], Lars Bernard[a]

christin.henzen@tu-dresden.de, stefano.della_chiesa@tu-dresden.de, lars.bernard@tu-dresden.de

[a]Chair of Geoinformatics, Technische Universität Dresden, Germany

Abstract. Most research activities in Earth System Sciences (ESS) are data-driven. There is a growing need to establish innovative, cross-cutting data management and data analysis methods in ESS to support the collaboration of interdisciplinary research building on heterogeneous sources. Data management plans (DMPs) are structured documents that outline data handling and include for instance agreements on roles, specifications of data products, and definition of workflows. However, the structure of existing DMP templates is mostly designed for funder's requirements and consequently address only the broad and interdisciplinary research community. Thus, these templates do lack (1) guidance on how to structure domain-specific information in a DMP – by providing domain-specific profiles, e.g. to harmonize the structure and improve the comprehensibility of DMP instances and (2) (linking into) tools enabling efficient management and reuse of information / sections of DMP instances. Therefore, we provide a concept of future DMP templates and address geo-domain-specific requirements, and the integration of DMPs into research data infrastructures. We recommend integrating structured provenance and quality information, using established concepts, and define a pathway to link tools into research data infrastructures, such that they foster automation of data management workflows and data reuse.

Keywords: data management plans, research data infrastructures, research data management, quality information, provenance

1 Introduction

Most Earth System Science (ESS) research projects are data-driven and produce datasets as main results. This contribution is framed within the GeoKur project[1], which aims to support the curation and quality assurance of ESS data. The developments include approaches to support Research Data Management (RDM) for the discovery, reuse, and collaboration, cross-domain research including heterogeneous data sources. Data management plans (DMPs) are formal and structured documents that capture all relevant information along the data lifecycle and typically a DMP prescribes measures for data description, data archiving, data access, data use and data interpretation. A DMP instance implements a certain DMP template in order to address specific funders e.g. the EU Framework Programme for Research and Innovation Horizon 2020[2], the German Science Foundation (DFG)[3] or the Federal Ministry of Education and Research (BMBF[4] and eventually address domain-specific requirements[5] (Figure 1).

[1] https://geokur.geo.tu-dresden.de/

[2] https://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx (downloadable template), https://ec.europa.eu/programmes/horizon2020//en/what-horizon-2020, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm (guideline)

[3] https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf

[4] https://www.cms.hu-berlin.de/de/dl/dataman/muster-dmp-bmbf

[5] https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf

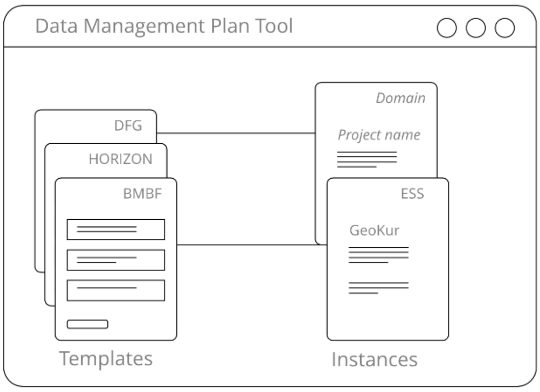


Figure 1: Data Management Plan templates and instances managed in a DMP tool.

DMPs are evolving and several communities [6] are addressing some of the challenges in the provision of dynamic DMPs as living documents, used collaboratively along the data life cycle, stored in catalogues [7]. This includes requirements on being machine-actionable (Miksa et al. 2019), domain specific through the Data Domain Protocols (DDP) [8] and ultimately making DMPs findable, accessible, interoperable, and reusable (FAIR) (cp. Wilkinson 2016). All of these challenges, in particular making DMPs FAIR, machine-actionable and domain-specific are closely linked and will improve transparency, reuse of information and provide interoperability within a data infrastructure e.g. using sections of DMP instances to fill dataset metadata (Davidson et al. 2019).

Despite all the efforts, RDM still lacks clear guidance and community specific and adapted DMP templates that meet domain-specific requirements (cp. Burnette et al. 2016). Finally, in this paper we address domain-specific aspects, which should improve future DMPs for ESS.

2 Include ESS specific content in the DMP templates

Data management has become an integral part of ESS research projects fostering efficient, collaborative, long-term data-driven research. Therefore, ensuring acquisition of fitness for use quality assured data and metadata is of paramount importance in the data life cycle. DMPs foster data and metadata management from the project's beginning and during the project execution. To establish a common understanding of data management from the beginning on, researchers need to ensure the provision of detailed data information by (1) choosing a proper metadata schema for the data description, and by (2) using the schema to annotate workflows (e.g. scripts), and to structure and export (automatically) generated metadata.

Here, we focus on the linkage of DMP templates and metadata, in particular provenance and quality information being highly relevant for ESS data. Provenance information is part of the metadata that describes the origin or history of data and foster the discovery, evaluation and reproducibility of research output and the data sources that generated it (Moreau et al. 2011, Simmhan et al. 2005).

Quality information is a fundamental characteristic of data, providing the necessary information to determine the fitness for use. It plays a major role in ESS, because earth observations, e.g. physical or chemical processes, are measured once, and the measured data is reused multiple times (Yang et al. 2012). Thus, a meaningful description of the data's quality is essential.

Within DMP templates, structured quality information, in particular, combined with provenance information, support the evaluation of datasets or complex workflows and the identification of workflow's core impacts or the relative weight/significance of some of its components.

By linking to existing information, we can reduce redundancy in the DMP instances while improving the detailed description of quality. Therefore, we recommend reusing the structure of related metadata schemas within ESS-specific DMP templates.

2.1 Include structured provenance information

To assess the fitness for use of a dataset and to provide all the elements for the reproducibility of a research output, tracing the origin of a result, recording the methods applied and data sources used, are fundamental. Thus, we argue including structured provenance fields in future DMP templates

1) to support the concept of FAIR data and FAIR workflows being linked to DMPs on a conceptual level, e.g. providing FAIR descriptions for all

[6] https://www.rd-alliance.org/groups/dmp-common-standards-wg, https://www.rd-alliance.org/groups/discipline-specific-guidance-data-management-plans-wg, https://rdmorganiser.github.io , https://www.gfbio.org/de/services, https://confluence.egi.eu/display/EOSC/Data+Management+Plan+Tool

[7] https://libereurope.eu/working-group/research-data-management/plans/

[8] https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf

elements of a workflow – datasets and geooperators – and on a modelling level, e.g. supporting the model-driven development of FAIR workflows[9] – generate executable code (fragments) from DMP's workflow description,

2) to facilitate the (automated) user-friendly visualization of provenance information – included in the DMP or as link to a visualization tool, e.g. showing interactive provenance graphs to better understand complex workflows (cp. Wagner et al. 2020, Henzen et al. 2016),
3) to support the automated integration of provenance information by bi-directionally linking of i) DMP tools, ii) workflow tools, used to manage and update workflows, and iii) metadata catalogues, used to store dataset's metadata including provenance information (see Section 3).

In ESS, several schemas and profiles to describe provenance information are well-known and well-used. Here, we recommend using ISO19115-1:2014 (ISO19115-1:2014) or PROV-O (Moreau and Missier 2013, Lebo et al. 2013) – both being compatible to frequently used and established research data management tools, such as metadata catalogues or annotation scripts. ISO19115-1:2014 describes provenance information of geoprocessing workflows on dataset level. PROV-O, as a cross-domain, extendable, linked data concept, captures provenance information on several levels of details, e.g. on dataset level as used in the GeoDCAT application profile (GeoDCAT-AP 2016). Thus, PROV-O can be easily linked to thematic or quality vocabularies.

2.2 Include structured quality information or data

When integrating and harmonizing heterogeneous data sources, like in most ESS projects, several quality aspects, such as uncertainties in the datasets, play a major role. Hence, quality information are required to evaluate and discover suitable inputs as well as to curate and assure derived outputs. Thus, we assume future DMP templates to

- link to quality registers. Establishing quality registers to manage and publish descriptions of quality dimensions and measures is key to make quality descriptions comparable, reusable, queryable and linkable, e.g. to domain-specific vocabularies. Several approaches for data quality vocabularies, such as DQV (Albertoni and Isaac 2021, Albertoni and Isaac 2016) can be used as a basis to structure quality information in registers.
- use cross-domain concepts for defining quality measures to enable comparison of data from several domains. For instance, essential variables (EVs) aim at having a common set of accurate and sustained measurements to ensure cross-domain usage (Patias et al. 2019, Lehmann et al. 2020). Therefore, climate[10], ocean[11] or biodiversity[12] addressing, i.a., cross-domain requirements on spatial and temporal resolution, uncertainty, and stability.
- link to ground truth data. In particular, remote sensing data are evaluated with ground truth datasets, and the results differ when ground truth data is changed. Linking to ground truth will improve the detailed evaluation of quality parameters and the reusability of the data.
- provide and link to quality datasets, showing geo-located quality information, facilitating the detailed evaluation of certain regions and the automated usage of quality information, e.g. as input dataset in modelling tools.
- describe sources of quality information to underpin the reliability of the given information. In ESS, quality information can be derived from the data, from metadata, from publications or from other sources. In particular, the information of quality in publication text have a summarized and qualitative attribute, which inevitably lead to information loss.

3 Integrate DMP Tools in Research Data Infrastructures

The components of a research data infrastructure (Figure 2) aim to systematically manage research data and support researchers during all phases of the data life cycle.

[9] https://fair-workflows.github.io/project.html
[10] https://gcos.wmo.int/en/essential-climate-variables
[11] https://www.goosocean.org/index.php?option=com_content&view=article&id=170&Itemid=114
[12] https://geobon.org/ebvs/what-are-ebvs/

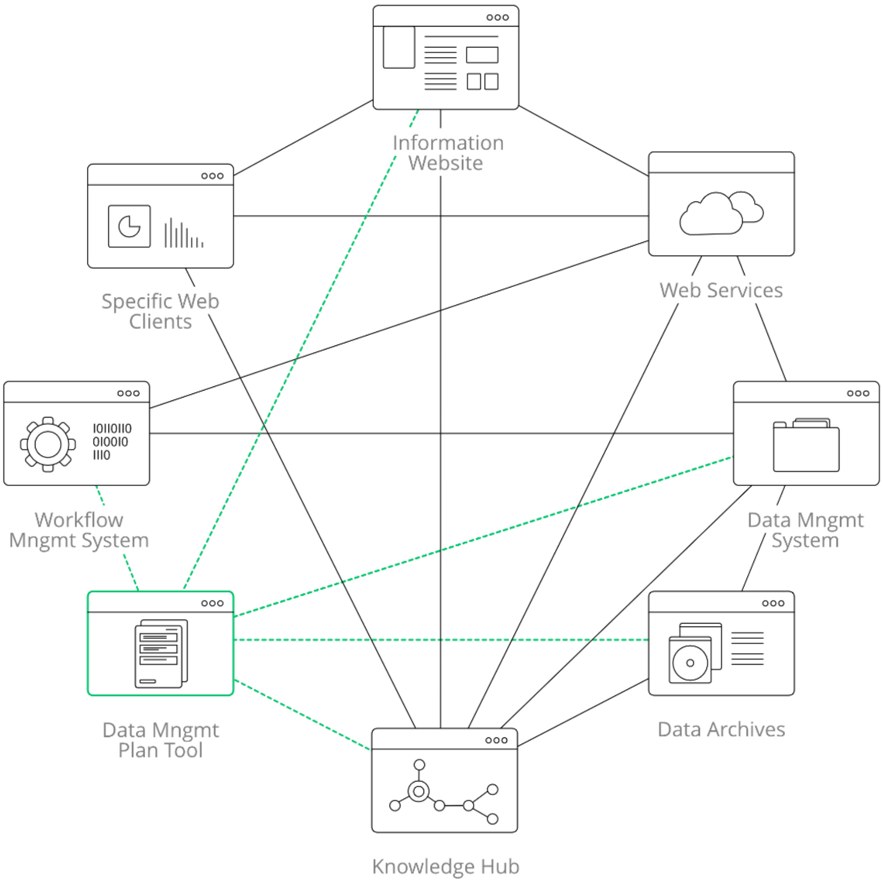


Figure 2: Components of a research data infrastructure with selected links.

The Web components include a Web site, which describe the project contents, specific Web clients developed for specific purposes such as. for data or provenance visualization, and Web services, e.g. for data publication and access (Figure 2, top). These components facilitate the provision of (project related) information and services..

The infrastructures include data management systems (e.g. catalogues like CKAN[13]), which supports efficient collaborative data management and are linked to Web services (for data publication) and to data archives for long-term storage (e.g. PANGAEA or organizational archiving systems, like OpARA[14]) (Figure 2, lower right).

Further, a knowledge hub enables managing and linking of domain knowledge, like thematic ontologies, vocabularies, and registries. Thus, data management systems and archives link to the hub to support meaningful descriptions for the provided data facilitating discovery and comprehensiveness. Knowledge hubs can be implemented as graph database or triple store to provide access to the knowledge as linked data and enable linking to external knowledge hubs (Figure 2, bottom).

Workflow management systems (e.g. Taverna or Kepler) can complement research data infrastructures by enabling the collaborated development of complex scientific workflows (cp. Wolstencroft et al. 2013, Ludäscher et al. 2006) (Figure 2, middle left). These tools can use data by linking to the data-related components, e.g. Web Services. However, workflow scripts, which are not managed in specific systems, can be linked to DMS, e.g. the package CKANR[15] enables using published data from the CKAN in R scripts.

All components[16] or part of it can be either configured for a certain research project on dedicated servers, managed by the related research organization to be used for all projects of the organization, or used as online instance for several organizations and projects.

[13] https://ckan.org/
[14] https://opara.zih.tu-dresden.de/xmlui/
[15] https://cran.r-project.org/web/packages/ckanr/index.html

[16] To simplify, we omitted project management and software development tools, like version control and issue tracker, which can be included as well.

However, in smaller projects or those with a specific focus, only selected components can be used.

To guide the researcher resp. data manager through the creation of DMP instances, several tools have been developed (Jones et al. 2020). DMP tools provide essentially a questionnaire on the data life cycle management plan, and, several different questionnaire (related to the templates) were tailored to specific requirements.

Frequently used online DMP tools[17] are i.a.:

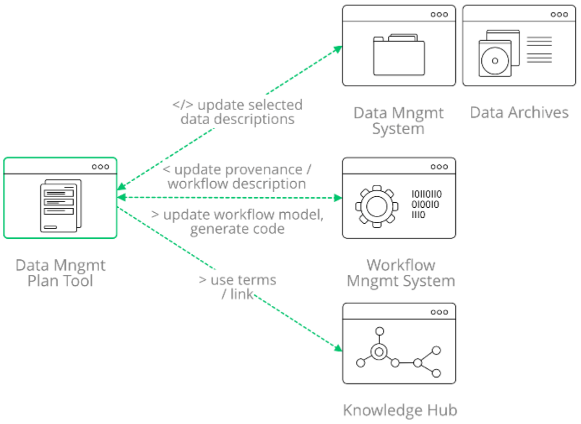
- RDMOrganiser[18] by several research institutions[19]
- Argos[20] by OpenAIRE
- DMPonline/DMPRoadmap[21] by Data Curation Centre
- GFBio Data Management Plan Tool[22] by gfbio
- MOSES Data Management Plan[23] by GFZ

Here, we see DMP tools as one of the core components of future research data infrastructures, being linked to several other tools (Figure 2, green box).

3.1 Foster automation by linking DMP tools with research data infrastructure components

Automation is core to support an efficient research data management, reducing the efforts to capture and update metadata, and to enable DMPs as living document. To improve automation, linking DMP tools to research data infrastructures is essential. We assume future DMP tools should (Figure 3):

- link to metadata catalogues bi-directionally to enable synchronization of relevant aspects included in both systems 1) when updated data and metadata are published in the metadata catalogue or 2) when concepts are updated, e.g. using different data or updating/replacing (parts of) geoprocessing workflows, in the DMP instance via DMP tool. Some DMP tools already provide export functionality of metadata for data catalogues, but are still missing the backlink from data catalogue to DMP tool resp. do lack metadata validation (before publishing to the linked catalogues).
- link to workflow management systems, e.g. to better update provenance information for geospatial workflows and to support model-driven workflow development starting from the structured DMP description in the DMP instance to generate code snippets. Bidirectional linking of workflow management system and DMP tool enables users to decide, whether they start with a general and formal description of provided workflows in the DMP instance, then generate and fill code, or first, model / implement a workflow in the workflow management system resp. use existing workflow models and then generate descriptions for the DMP.
- be linked to DMP catalogues to enable discovery and storage of DMP instances to make them FAIR. The Web tools DMPonline[24] and Catalogue of the LIBER RDM Working Group[25] already provide discovery of DMPs, but do lack proper filter mechanism to support users in finding suitable DMP instances.
- link to existing registries, vocabularies, or ontologies, e.g. to enable discovery and comparison of DMP instances by using these links as search filters for DMP instances and making DMP instances comparable. Managing DMPs as linked data and linking to existing information enables semantic querying and reusing of the links, e.g. to be exported for data description in catalogues.



</> update selected data descriptions
Data Mngmt System
Data Archives
< update provenance / workflow description
> update workflow model, generate code
Data Mngmt Plan Tool
Workflow Mngmt System
> use terms / link
Knowledge Hub

Figure 3: DMP tool and recommended future linking (arrows show direction of linking).

[17] Further lists of tools are available here: https://www.forschungsdaten.org/index.php/Data_Management_Pl%C3%A4ne and here: https://activedmps.org/.

[18] https://rdmorganiser.github.io/kooperationen

[19] Leibniz Institute for Astrophysics Potsdam (AIP), Potsdam University of Applied Sciences (FHP), Karlsruhe Institute of Technology Library (KIT)

[20] https://argos.openaire.eu/splash/

[21] https://dmponline.dcc.ac.uk/

[22] https://www.gfbio.org/plan

[23] https://moses-dmp.gfz-potsdam.de/

[24] https://dmponline.dcc.ac.uk/public_plans

[25] https://libereurope.eu/working-group/research-data-management/plans/

4 Conclusion

Data Management Plans and related DMP tools are core for data-driven research projects. Linking DMP tools as core components of future research data infrastructures, to several other tools facilitates reusing of information and reduces efforts in metadata, data and DMP management.

Within the GeoKur project, we develop concepts, best practices, and develop RDM components supporting curation and quality assurance of ESS data, e.g. a metadata extraction tool (Wagner et al. 2021) or a Geo-Dashboard visualizing provenance and quality information (Figgemeier et al. 2021). We use and adapt open-source tools, like the Fuseki triple store as knowledge hub for data quality, provenance and land use ontologies, the CKAN to implement a GeoDCAT profile, and the RDMO tool to be extended and linked to the CKAN. Our requirement analysis for the components includes high efforts in gathering meta information for potential analysis inputs, in particular addressing the gap of provided information and information needs for provenance and quality information. Taking GeoKur as underlying use case, we developed several recommendations on which and how meta information, in particular provenance and quality information, need to be included and structured within a DMP and how DMP tools should be linked within a research data infrastructure to facilitate efficient metadata management.

When following the provided recommendations the DMP tools need to be more flexible. They should store DMP instances format-independently to enable the mapping of DMP parts to several (dataset's or workflow) metadata formats or DMP output formats. Related concepts can be proposed on several levels of details: (1) by using (life sciences or natural sciences) Domain Data Protocols [26] (DDP) to cover domain-specific aspects for DMPs on a conceptual level, (2) implementing DMPs as linked data to enable semantic linking and querying, which could include integrating or linking to provenance (e.g. using PROV-O) or quality information (e.g. using DQV), and (3) by mapping to standards and languages, e.g. ISO metadata standard ISO 19115 or Business Process Modeling Language to provide APIs for coupling data, workflow, and DMP tools. Thus, guiding researchers on how to structure domain-specific information in a DMP, and linking DMP tools to other RDM components, support data management practices with machine-actionable richness, creating ultimately, living, dynamic DMPs with a greater added value for all the stakeholder.

Acknowledgement

The authors thank Lukas Egli, Julia Fischer, and Juliane Groth (Helmholtz Centre for Environmental Research for their inputs regarding information needs and DMP examples.

The presented recommendations have been developed in the project GeoKur. GeoKur is granted by the BMBF (Federal Ministry of Education and Research) under number 16QK04A.

References

Albertoni, R., Antoine, I. (2020): Introducing the Data Quality Vocabulary. Semantic Web, Vol. 12, No. 1, pp. 81-97, DOI: 10.3233/SW-200382, 2021.

Albertoni, R., Isaac, A. (Eds.) (2016) Data on the Web Best Practices: Data Quality Vocabulary. https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/.

Burnette, M. H., Williams, S.C., Imker, H. J. (2016): From Plan to Action: Successful Data Management Plan Implementation in a Multidisciplinary Project. Journal of eScience Librarianship 2016;5(1): e1101. https://doi.org/10.7191/jeslib.2016.1101, 2016.

Davidson, J., Engelhardt, C., Proudman, V., Stoy, L., Whyte, A. (2019): D3.1 FAIR Policy Landscape Analysis. URL: https://zenodo.org/record/3558173#.YGWC750zZ3h, 2019.

Figgemeier, H., Henzen, C., Rümmler, A. (2021): A Geo-Dashboard Concept for the Interactively Linked Visualization of Provenance and Data Quality for Geospatial Datasets. AGILE Conference 2021, virtually.

GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe. URL: https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/geodcat-application-profile-data-portals-europe/distribution/geodcat-ap-101-docx, 2016.

[26] Science Europe Guidance Document - Presenting a Framework for Discipline-specific Research Data Management, https://www.scienceeurope.org/media/nsxdyvqn/se_guidance_document_rdmps.pdf

Henzen, C., Mäs, S., Zander, F., Schroeder, M., Bernard, L. (2016): Representing Research Collaborations and Linking Scientific Project Results in Spatial Data Infrastructures by Provenance Information. 19th AGILE Conference on Geographic Information Science, Helsinki, 2016.

ISO 19115-1 International Standard on Geographic information - Metadata - Part 1: Fundamentals. 1st Edition, 2014.

Jones, S., Pergl, R.; Hooft, R., Miksa, T., Samors, R., Ungvari, J. et al. (2020): Data Management Planning: How Requirements and Solutions are Beginning to Converge. In Data Intelligence 2 (1-2), pp. 208–219. DOI: 10.1162/dint_a_00043, 2020.

Lebo, T., Sahoo, S., McGuinness, D. (2013): PROV-O: The PROV Ontology. W3C Recommendation. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/, 2013.

Lehmann, A., Masò, J., Nativi, S., Giuliani, G. (2020): Towards integrated essential variables for sustainability. International Journal of Digital Earth, Vol. 13, pp. 158-165, https://doi.org/10.1080/17538947.2019.1636490, 2020.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y. (2006): Scientific workflow management and the Kepler system. In Concurrency Computation Practice and Experience. 18 (10), pp. 1039–1065. DOI: 10.1002/cpe.994, 2006.

Miksa, T., Simms, S., Mietchen, D., Jones, S. (2019): Ten Principles for maDMPs. PLoS Comput Biol 15(3): e1006750, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006750, 2019.

Moreau, L., Clifford, B., Freire, J., Gil, Y., Groth, P., Futrelle, J., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Simmhan, Y., Stephan, E., Van den Bussche, J (2011): The Open Provenance Model - Core Specification (v1.1); Future Generation Computer Systems, 2011.

Moreau, L., Missier, P. (2013): PROV-DM: The PROV Data Model. W3C Recommendation. URL: http://www.w3.org/TR/2013/REC-prov-dm-20130430/, 2013.

Patias, P., Verde, N., Tassopoulou, M., Georgiadis, C., Kaimaris, D. (2019): Essential variables: describing the context, progress, and opportunities for the remote sensing community. Proc. Vol. 11174, Seventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2019); 111740C, https://doi.org/10.1117/12.2533604, 2019.

Simmhan, Y., Plale, B., Gannon, D. (2005): A Survey of Data Provenance in e-science, SIGMOD record; Vol. 34, pp. 31-36, 2005.

Wagner, M., Henzen, C., Müller-Pfefferkorn, R. (2021): A Research Data Infrastructure Component for the Automated Metadata and Data Quality Extraction to Foster the Provision of FAIR Data in Earth System Sciences. AGILE Conference 2021, virtually.

Wagner, M., Rümmler, A., Henzen, C., Mäs, S., Müller-Pfefferkorn, R., Bernard, L. (2020): A User-Centred Approach to Foster Research Data Management in Earth System Sciences by Providing User-Friendly Visualizations for Automated Generated Data and Process Metadata. 16th Research Data Alliance (RDA) Plenary "Knowledge Ecology", Costa Rica / virtually, 2020.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G, Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R. , Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, Vol. 3, AN: 160018, 2016.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A.N., Balcazar Vargas, M.P., Sufi, S., Goble, C. (2013): The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. In Nucleic acids research 41 (Web Server issue), W557-61. DOI: 10.1093/nar/gkt328.