IMPLEMENTATION ARTICLE

Related to other papers in this special issue	3 (p30); 11 (p108)
Addressing FAIR principles	F1, F2, F3, F4, A1, R1

Making Data and Workflows Findable for Machines

Tobias Weigel[1†], Ulrich Schwardmann[2], Jens Klump[3], Sofiane Bendoukha[1] & Robert Quick[4]

[1]Deutsches Klimarechenzentrum, Bundesstrasse 45a, Hamburg 20146, Germany

[2]Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen, Am Faßberg 11, 37077 Göttingen, Germany

[3]CSIRO, Kensington, WA 6151, Canberra, Australia

[4]Indiana University Bloomington, Bloomington, IN 47405, USA

Keywords: Findability; Workflows; Automation; FAIR data; Data infrastructures; Data services

Citation: T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. Data Intelligence 2(2020), 40–46. doi: 10.1162/dint_a_00026

ABSTRACT

Research data currently face a huge increase of data objects with an increasing variety of types (data types, formats) and variety of workflows by which objects need to be managed across their lifecycle by data infrastructures. Researchers desire to shorten the workflows from data generation to analysis and publication, and the full workflow needs to become transparent to multiple stakeholders, including research administrators and funders. This poses challenges for research infrastructures and user-oriented data services in terms of not only making data and workflows findable, accessible, interoperable and reusable, but also doing so in a way that leverages machine support for better efficiency. One primary need to be addressed is that of findability, and achieving better findability has benefits for other aspects of data and workflow management. In this article, we describe how machine capabilities can be extended to make workflows more findable, in particular by leveraging the Digital Object Architecture, common object operations and machine learning techniques.

[†] Corresponding author: Tobias Weigel (E-mail: weigel@dkrz.de, ORCID: 0000-0002-4040-0215).

1. INTRODUCTION

In several scientific disciplines, the number, size and variety of objects to be managed are growing. Examples of particular interest to the challenges discussed in this article include climate modeling [1], geophysics [2], and "-omics" [3]. The supporting data infrastructures and services are challenged to offer adequate solutions, and are looking toward increased automation in their processes to cope with the needs. Aspects of automation are intrinsic to the FAIR vision [4]. This article highlights the steps that are required to automatically identify objects, associate them with metadata, and make both data and the processes that generated them more findable. Persistent identifiers, machine processes with autonomous decision-making capability, and machine-actionable metadata are critical elements for practical solutions.

The motivation is given through the increased interest by researchers and funders in making not only those data available that underpin analysis in scientific publications, but also give insight into the generative history of these data while they were generated, processed, analyzed and eventually published. Readers wish to investigate the provenance of data underlying publications, gaining access to context information on data in the provenance graph and on workflows or individual data processing steps. In this article, we investigate how such information can be aggregated and leveraged to improve the general findability of data and workflows that produce them, improving the quality of information that search catalogs such as B2FIND[1] or the CSIRO Data Access Portal[2] can depend upon. The potentially following step—to enable machines to find resources automatically as part of orchestration—will only be touched marginally. Concerning aggregation for findability, the article highlights key requirements and elements of possible solutions that can inform future work.

Researchers who work with data are also interested in making their workflows more efficient, shortening the time from data production to analysis, but also short-cutting workflows, for example, when using in-situ visualization in a High-Performance Computing (HPC) workflow to detect errors already during a computing run and restarting the process quickly with modified parameters. Another important usage trend is the motivation of users to work with data at higher levels of abstraction. Researchers are increasingly relying on tools such as Jupyter notebooks and standard software libraries to deal with issues of data access and management, giving rise to the wider adoption of Virtual Research Environments (VREs; e.g., [5, 6]). It is much more efficient to let them focus on the scientific questions of data analysis, and reduce the amount of resources they spend on data management and access. This is part of a larger cultural change, which has wide impact on the evolution of data services, and improving findability is a key concern.

A key capability necessary to support future scenarios is, therefore, support at data infrastructure level for better automation of the processes dealing with data and workflows. Out of the many possible facets related to this challenge that could be derived from the FAIR principles, in this article, we focus on the automation of findability (principles F1-F3), emphasizing that identifiers are a foundational element from

[1] https://b2find.eudat.eu.

[2] https://data.csiro.au/dap.

which the other principles must follow [7]. A key question is: How can automated processes help to make more data and workflows findable, particularly from early research workflow stages? In this article, we understand an automated process as one that is capable of limited, autonomous decision-making. This is driven by rule systems specified by humans, but could also, in a later evolution, be replaced by means of machine learning.

2. ESSENTIAL REQUIREMENTS FOR AUTOMATING DATA AND WORKFLOW FINDABILITY FOR MACHINES

To support machine-actionable processes in data infrastructures and VREs, objects (including data and workflows, but possibly also other artefacts) need to be persistently identifiable independent from location (F1) [7]. This is the primary prerequisite for any other benefits. An important constraint is that the future preservation status of an object is likely unclear at early stages, but identification is nonetheless required. This shapes the choice of persistent identifier (PID) systems to employ.

Moreover, offering elemental operations on objects independent from location can support the needs of data management processes within data infrastructures well. We can define two levels of such operations. The first level includes “create, read, update and delete” (CRUD) operations on single objects and collections, as well as directly related support operations such as retrieving object metadata, which is critical to facilitate findability. The second level consists of more complex operations such as creating a replica (physically copy an object, update metadata to describe replica) or creating a success or version (create a new object, describe relation with predecessor in metadata).

In order to record a fundamental level of provenance from data processing in VREs, persistent identification must be complemented by additional metadata (F2). Following the PROV model [8], the first action is to record the link between input and output data using *wasDerivedFrom* relations between entities. Extensions may then add links to the processing script or workflow (activity) and links to the executing user (agent). Finally, to enable automated processes to not only use such information (which can be ensured by relying on PROV-compatible encodings) but also discover it in the first place, the methods by which to access it must be well-defined (A1). This does not necessarily mean that the information has to be centrally stored. It could also be federated, but the mechanism has to be straightforward to use by automated processes, and the use of the identifier to access metadata according to A1 is a primary prerequisite.

One particular aspect for data processing is that workflows may be defined by users, but will be best used by automated processes as (Web) services, i.e., they must have a machine-interpretable service description, and standardized interfaces, and give back unambiguous information about execution results. A mechanism that makes such information readily available for automated agents is required.

One well-known established approach for addressing concerns of reproducibility, automation and provenance, in particular, are scientific workflow systems (e.g., [9]). These have seen larger adoption in the “-omics” research area, but are less adopted for climate or geophysics data processing scenarios, in contrast to the adoption of interactive Python via Jupyter notebooks. One contributing factor may be that interactive notebooks support both repetitive tasks and exploratory work modes. Exploration is, for instance, a major factor for geophysics use cases in the Scientific Software Solution Centre (SSSC) [2], where the set of possible input data is relatively small, but there is a large variation in algorithms to evaluate. Cases like this make the Jupyter notebook particularly attractive. For HPC environments and data infrastructures, integration with fairly heavy-weight workflow systems poses a huge challenge that is often refrained from due to the large investment required. Nonetheless, the controlled environment of a workflow system has a certain appeal for automating metadata capture in the background (F2, R1), yet in view of the tremendous user adoption of interactive notebooks, a solution should potentially be suitable for both approaches.

Finally, a few other general requirements will be critical to ensure practical success of any solution. It should be resilient against operational incidents, such as temporary server outages, and either recover from them without human intervention or fully fail over, ensuring continuous operation. While this may seem a generally good requirement for any technical system, it is even more critically so if the system is built on automated processes capable of limited autonomy.

3. ELEMENTS OF A POSSIBLE SOLUTION

A consistent implementation that improves findability for machines can be broken down into several parts, following the research workflow. Ultimately, findability depends on trustworthy, gapless metadata, and improving the processes in the workflow contributes to better findability at the consumer's end. One important constraint is that even a full solution will still be semi-automatic, i.e., a human user will still need to be involved as the ultimate referee, contributor of some metadata elements only a human can define, and to ensure overall quality control.

The persistent identification of objects and reciprocal association (F3) with machine-interpretable metadata can be facilitated by employing the Digital Object Architecture (DOA) [10], PID Kernel Information [11] and Data Type Registries [12, 13], also following the model of FAIR Digital Objects [14]. In the following, we describe along a generalized workflow how these elements can be combined to address the requirements. A key aspect is that the PID is seen as the primary anchor or "entry point" for any agent interacting with an object, making availability and proper maintenance of PIDs a necessary precondition.

As a first step, any object should receive a PID. At the earliest workflow stages, descriptive metadata are likely not available yet, and the long-term preservation status of an object is also unclear. Therefore, a PID system that does not mandate specific metadata elements and does not enforce strong policies, such as the Handle System, is a good fit. Even more important than the choice of PID system is that, in order to support later operations and autonomous handling of objects, the PID should be embedded in the object. For instance, if the object is a file with a format supporting embedded metadata, the PID should be included.

While descriptive metadata may not be available, support for generalized CRUD operations requires essential structural and administrative metadata to be captured, stored and made available for requestors.

Metadata capture must be highly automated and reliable, both in terms of technical reliability and ensured metadata quality. This requires an approach that may be very different from established procedures. For example, in the case of adoption by the Earth System Grid Federation (ESGF[3]) for climate data, it became clear very early that technical solutions must be embedded in processes agreed with all stakeholders (users, project and data managers, infrastructure providers and administrators), and that defining and establishing these processes is a prerequisite for subsequent technical development. This leads to a general observation that, in particular, the quality of metadata may be controlled by technical means, but high quality can only be achieved if the processes are supported by all stakeholders.

Metadata delivery must work with low latency and in highly standardized, machine-interpretable encodings. While originally addressed toward encoding provenance, the concept of PID Kernel Information and its underlying principles can fulfill these additional requirements also to support metadata required for CRUD operations. At later workflow stages of data publication and wide sharing, PID Kernel Information is unsuitable to feed search catalogs with meaningful descriptive metadata, and must be complemented with other sources. Here, it may be possible to leverage existing workflows driven by the needs for descriptive metadata for purposes of archival and credit-giving.

Common operations on objects may best be implemented according to a comprehensive specification. In the DOA framework, a Digital Object Interface Protocol (DOIP) has been defined to additionally incorporate such operation specifications. Implementing such a protocol on top of not just a single repository, but as part of a service-oriented architecture may be more difficult, since it is likely that any single operation is offered by multiple middleware services, and that execution of an operation may have side effects or require actions to be taken by other services. A practical obstacle for implementation is the required compatibility with existing architectural components, protocols and interfaces (e.g., REST), and fitness for use within distributed systems, where no single control point may exist to coordinate the execution of operations.

Services and VREs for data analysis and processing such as the Scientific Software Solution Centre (SSSC) [2] or the ENES Climate Analytics Service (ECAS) [15] present a unique opportunity to implement a solution at small scale in a relatively closed environment, since their supported workflows are a smaller but representative subset of the more general research data workflow, and the central control over the VRE workflows makes implementation easier compared to distributed middleware in larger data infrastructures. Implementing automated PID assignment, metadata generation and provenance capture and elemental object operations in a VRE may easily demonstrate improvements to downstream findability that can inform decisions on implementations in larger infrastructures.

[3] https://esgf.llnl.gov.

4. EXTENDING CAPABILITIES WITH MACHINE LEARNING

We will briefly highlight two opportunities for a solution to employ machine learning techniques, concerning classification for search catalogs (F4) and building recommender systems. While the components mentioned so far can improve metadata acquisition, it is likely that gaps will remain that also cannot be covered through increased human intervention. Machine learning may help by classifying artefacts based on incomplete information, possibly also using unstructured sources such as log files from executing computing jobs or running processing tools. This may work particularly well if a VRE or scientific workflow management system is used, but may also work well in the back of common HPC jobs.

The most important constraint is that the result of such classification by machine learning algorithms will bear an intrinsic uncertainty. It should therefore not be a full alternative to metadata acquisition, particularly in view of the level of precision required for data preservation, but it can fuel search catalogs, as a level of uncertainty may be tolerable. Information both out of metadata and algorithmic classification may then be used to power recommender systems that enhance search catalog capabilities [16]. They may recommend, for example, input data sets or workflows for reuse to VRE users, and thus contribute to improved findability from the consumer's end.

5. OUTLOOK AND CONCLUSIONS

We have touched upon important requirements and key elements of a solution to improve findability of data and workflows by leveraging automation capabilities during the research workflow. Future work on the topic may derive more concrete recommendations and build demonstrators. The approach described, motivated by requirements seen in several disciplines, hints at promising solutions that could emerge out of collaborative work across disciplinary infrastructures and service providers.

In the end, a decisive take on automation for findability can be of benefit to multiple stakeholders. Researchers producing data can spend less time on data management and documentation, researchers reusing data and workflows will have access to metadata on a wider range of objects, and research administrators and funders may benefit from deeper insight into the impact of data-generating workflows. The wider adoption of a solution may also benefit other aspects of FAIR, e.g., interoperability and reusability.

AUTHOR CONTRIBUTIONS

All authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript. Tobias Weigel (weigel@dkrz.de) has led the editorial process.

REFERENCES

[1] V. Balaji, K.E. Taylor, M. Juckes, B.N. Lawrence, P.J. Durack, M. Lautenschlager, … & D. Williams. Requirements for a global data infrastructure in support of CMIP6. Geoscientific Model Development 11 (9)(2018), 3659-3680. doi: 10.5194/gmd-11-3659-2018.

[2] G. Squire, M. Wu, C. Friedrich, L.A.I. Wyborn, J. Klump & R. Fraser. Scientific software solution centre for discovering, sharing and reusing research software. In: American Geophysical Union Fall Meeting 2018. Available at: https://agu.confex.com/agu/fm18/meetingapp.cgi/Paper/459873.

[3] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M.R. Crusoe, K. Peters & D. Schober. FAIR computational workflows. Data Intelligence 2(2020), 108–121. doi: 10.1162/dint_a_00033.

[4] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O. Bonino da Silva Santos & M.D. Wilkinson. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use, 37(1)(2017), 49-56. doi: 10.3233/ISU-170824.

[5] L.A.I. Wyborn, R. Fraser, B.J.K. Evans, C. Friedrich, J.F. Klump & D. Lescinsky. Building a generic virtual research environment framework for multiple earth and space science domains and a diversity of users. In: American Geophysical Union Fall Meeting 2017. Available at: https://agu.confex.com/agu/fm17/meetingapp.cgi/Paper/293857.

[6] M. Barker, S.D. Olabarriaga, N. Wilkins-Diehr, S. Gesing, D.S Katz, S. Shahand, … & A. Costa. The global impact of science gateways, virtual research environments and virtual laboratories. Future Generation Computer Systems 95(2019), 240-248. doi: 10.1016/j.future.2018.12.026.

[7] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(2020), 30–39. doi: 10.1162/dint_a_00025.

[8] L. Moreau, P. Missier (eds.). K. Belhajjame, R. B’Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, … & C. Tilmes. PROV-DM: The PROV Data Model. W3C Recommendation REC-prov-dm-20130430. World Wide Web Consortium (2013). Available at: https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[9] I. Taylor, E. Deelman, D.B. Gannon & M. Shields. (eds.). Workflows for e-Science—Scientific workflows for grids. London: Springer-Verlag, 2007. isbn: 978-1-84628-519-6.

[10] R. Kahn & R. Wilensky. A framework for distributed digital object services. International Journal on Digital Libraries 6 (2)(2006), 115-123. doi: 10.1007/s00799-005-0128-x.

[11] T. Weigel, B. Plale, M. Parsons, G. Zhou, Y. Luo, U. Schwardmann, … & K. Kurakawa. Recommendation on PID kernel information. Research Data Alliance (2018). doi: 10.15497/RDA00031.

[12] L. Lannom, D. Broeder & G. Manepalli. Data type registries working group output. doi: 10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458.

[13] U. Schwardmann. Automated schema extraction for PID information types. In: 2016 IEEE International Conference on Big Data, IEEE, 2016, pp. 3036-3044. doi: 10.1109/BigData.2016.7840957.

[14] The European Commission High Level Expert Group report: Turning FAIR into reality (2018). doi: 10.2777/1524.

[15] S. Bendoukha, T. Weigel, S. Fiore & A. D’Anca. ENES climate analytics service (ECAS). In: Proceedings of the European Geosciences Union General Assembly 2018. Available at: https://meetingorganizer.copernicus.org/EGU2018/EGU2018-12549.pdf.

[16] A. Devaraju & S. Berkovsky. A hybrid recommendation approach for open research datasets. In: Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, ACM, pp. 207–211. doi: 10.1145/3209219.3209250.