Data Provenance

Barbara Magagna[1](✉), Doron Goldfarb[1], Paul Martin[2], Malcolm Atkinson[3], Spiros Koulouzis[2], and Zhiming Zhao[2]

[1] Environment Agency Austria, Vienna, Austria
{barbara.bagagna,doron.goldfarb}@umweltbundesamt.at
[2] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
paulmartin.research@gmail.com, {s.koulouzis,z.zhao}@uva.nl
[3] University of Edinburgh, Edinburgh, UK
malcolm.atkinson@ed.ac.uk

Abstract. The provenance of research data is of critical importance to the reproducibility of and trust in scientific results. As research infrastructures provide more amalgamated datasets for researchers and more integrated facilities for processing and publishing data, the capture of provenance in a standard, machine-actionable form becomes especially important. Significant progress has already been made in providing standards and tools for provenance tracking, but the integration of these technologies into research infrastructure remains limited in many scientific domains. Further development and collaboration are required to provide frameworks for provenance capture that can be adopted by as widely as possible, facilitating interoperability as well as dataset reuse. In this chapter, we examine the current state of the art for provenance, and the current state of provenance capture in environmental and Earth science research infrastructures in Europe, as surveyed in the course of the ENVRIplus project. We describe a service developed for the upload, dissemination and application of provenance templates that can be used to generate standardised provenance traces from input data in accordance with current best practice and standards. The use of such a service by research infrastructure architects and researchers can expedite both the understanding and use of provenance technologies, and so drive the standard use of provenance capture technologies in future research infrastructure developments.

Keywords: Provenance · Scientific workflow management · Research data

1 Provenance in the Environmental Domain

One particularly sensitive issue in the context of environmental research data lifecycles is the provenance of offered data products. In order to allow scientific reuse, published research datasets need clear annotations detailing their genesis and any additional processing applied afterwards. This includes information about the methodology, instrumentation and software used in data acquisition, subsequent processing and preservation, covering all steps of the typical research data lifecycle. The collected information

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 208–225, 2020.
https://doi.org/10.1007/978-3-030-52829-4_12

should not only be targeted at a human audience but should also be machine-processable in order to support various forms of analysis for a variety of purposes such as the choice of suitable data sources or the assessment of patterns of re-use. The increasing availability of reusable, provenance-enabled datasets moreover requires researchers and engineers to consider their "second hand" provenance in addition to "first hand" locally-generated traces and the subsequent combination of these different streams for further reuse. This, even more, underscores the importance of consistent usage of dedicated and interoperable standards for representing provenance, especially in light of recent developments regarding requirements for to-be-published research data, such as the FAIR data principles.

While issues of reproducibility and scientific integrity of research results have traditionally been a central concern for any scientific domain, current environmental developments on global scales often trigger controversies about the underlying cause-effect scenarios. This sometimes even leads to mutual accusations of politically or ideologically driven manipulations of data and resulting scientific evidence. Such a strong political relevance of contemporary environmental research data thus underscores the importance of adequate protocols allowing to trace the respective results back to their origin, acting as evidence for their soundness.

Given the scenarios sketched above, there is a clear and increased need for environmental research infrastructures to develop and maintain well-established provenance generation, provision and tracking infrastructure which is interoperable across the overall landscape of involved domains. Unfortunately, this requirement is hampered by the great heterogeneity of approaches to environmental research and the resulting spectrum of environmental research infrastructures, characterised by a wide variety of objects of interest, applied acquisition and overall research methodologies. Services aiming to cater the needs of the individual research workflows would thus either have to be very specific, or as generic as possible.

In this chapter, we survey the state of the art of provenance gathering and visualisation technologies and standards and describe how we addressed the heterogeneity of research infrastructure in the context of the ENVRIplus project[1], which was charged with the development of generic common services to assist the development and interoperability for environmental and Earth science research infrastructures (RIs) in Europe. We review some of the requirements of RIs regarding research data and data process provenance, and we describe a system for producing, sharing and instantiating provenance templates online, which we believe can help RI architects and engineers, as well as general researchers, to produce better-standardised provenance traces that can be interpreted in a broad range of different contexts.

2 State of the Art

Although there exist several provenance models used in specific settings promoted by different international initiatives, the main basic standard widely-used and referred to is the W3C's PROV recommendation[2], which evolved from the Open Provenance Model

[1] https://www.envriplus.eu/.

[2] https://www.w3.org/TR/prov-overview/.

(OPM). After three international workshops [1] on provenance standardisation, OPM was developed in 2010 and subsequently adopted by many workflow systems. PROV is very much influenced by OPM and was released as a standard by the W3C Provenance Working Group in April 2013. The W3C PROV Recommendation [2] consists of some constituent standards including for PROV XML[3] and PROV as an ontology for RDF-based data (PROV-O)[4].

The essential elements (see Fig. 1) of the PROV ontology are called Starting Point Terms and consist of three primary classes with unique and mandatory identifiers and nine properties to describe the relations between the classes. The three classes are prov:Entity which is the central concept and represents resources, prov:Activity representing actions performed upon entities, and prov:Agent representing persons or machines who bears some form of responsibility for an activity. The most important relationships are: *used* (an activity used some artefact), *wasAssociatedWith* (an agent participated in some activity), *wasGeneratedBy* (an activity generated an entity), *wasDerivedFrom* (an entity was derived from another entity), *wasAttributedTo* (an entity was attributed to an agent), *actedOnBehalfOf* (an agent acted on behalf of another agent) and *wasInformedBy* (an activity used an entity produced by another activity). Expanded terms are used to specialise agents and entities and to introduce time validity descriptions for activities. *Qualified terms* are used to provide additional attributes of the binary relations introducing so-called qualified patterns. In this way it is possible to add for example the concept 'plan', an association class, to describe more in detail how an activity was carried out.

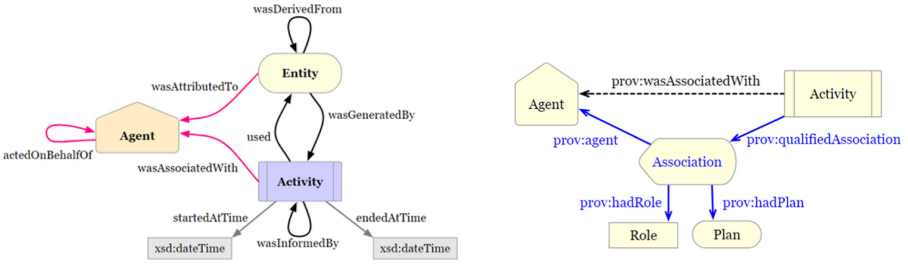

Fig. 1. PROV-O, starting point terms and qualified patterns.

W3C PROV has primarily been designed to describe retrospective provenance (r-prov) which refers to a-posteriori descriptions of provenance traces of data resources, i.e. provenance as an extended log of all the steps executed to generate the data entity. The concept of provenance can, however, also refer to tracing the genesis of workflows used for generating data, and moreover even to the a-priori description of such workflows, in which case it is called prospective provenance (p-prov) which can be considered to be a form of workflow description language.

In order to be able to represent workflow templates and workflow instances, Garijo and Gil extended PROV [3] to P-plan. OPMW [4], an extension of P-plan, PROV

[3] https://www.w3.org/TR/2013/NOTE-prov-xml-20130430/.
[4] https://www.w3.org/TR/2013/REC-prov-o-20130430/.

and OPM, is designed to represent prospective provenance of scientific workflows at a fine granularity. D-PROV [5] extended PROV with workflow structure, later being replaced by ProvONE[5], which can track all different types of provenances including the graph structure of the dataflow itself. S-PROV [6] is built upon the PROV and ProvONE models, helping the scientist to analyse the workflow at different levels of granularity and capturing runtime change. It is the underlying model for S-ProvFlow [7], a provenance framework for storage, access and discovery of data-intensive streaming lineage, used in the VERCE Earthquakes Simulation Portal[6] used by the EPOS[7] community. PROV-Wf [8], another specialisation of the W3C PROV-Data Model, allows the capture of both prospective and retrospective provenance but also supports domain-specific data provenance increasing the potential of provenance data analysis. Not all contemporary approaches to provenance are based on W3C PROV. One different approach is the WF4Ever Research Object description[8], conceived as self-contained units of knowledge, aggregating information about the generation workflow at a general level, not directly aligned but still mappable to W3C PROV. CERIF[9] (the Common European Research Information Format) is an entity-relationship model with temporal additions of the research domain used in the EPOS community. It supports the management of Research Information, including details on people, projects, organisations, publications and products. Instances of this representation provide some provenance information because of the time-stamped linking entities used to assets when certain relationships were formed. Nevertheless, CERIF needs further development of some provenance aspects related to the integration of causal-effect relationships among entities [9].

A variety of online tools are made available on dedicated websites[10] in order to support the use of the PROV standard in data management. Examples are the public provenance data storage based on ProvStore [10], the validator against the PROV standard as well as conversion services for various standard output formats. Dedicated libraries are in turn provided for including provenance functionality in local applications. ProvToolbox[11] is an example for a Java library providing different means for manipulating provenance descriptions and converting them between RDF, PROV-XML, PROV-N, and PROV-JSON encoding; comparable functionality for Python-based environments is provided by the PROV[12] Python package.

Starting an overview about technologies and approaches in use from the point of data acquisition, the first phase of the ENVRI RM research data life cycle (see figure page 14 of [11]), it is important to distinguish between manual scenarios and automated settings. Manual measurements and observations made using pen and paper would need to be transferred to spreadsheets or databases before existing provenance recording tools such as InSituTrac [12] could be applied. Shifting from pen-and-paper data collection

[5] http://tinyurl.com/ProvOne.
[6] http://portal.verce.eu.
[7] https://www.epos-eu.org/.
[8] http://wf4ever.github.io/ro/2016-01-28/.
[9] https://www.eurocris.org/cerif/main-features-cerif.
[10] https://openprovenance.org/ and https://provenance.ecs.soton.ac.uk/.
[11] https://github.com/lucmoreau/ProvToolbox.
[12] https://github.com/trungdong/prov.

to acquisition via handheld devices could therefore also improve provenance related aspects, such as demonstrated by the EcoProv [13] approach. Handheld applications such as developed in the Urbanopoly project [14] are moreover a potential platform for collecting provenance data in citizen science settings. As an example for tracing the genesis of manually curated scientific databases, the "copy-paste model" approach from [15] captures chains of insertions, deletions and imports from sources to a target database.

It is clear therefore that in many cases, data acquisition includes both manual and automatic aspects. In sample-based data collection, provenance capture should ideally start with the human sampling process and continue with the subsequent analysis taking place in laboratories. In this regard, the alignment of the ISO 19156 Sampling Features Schema with the W3C PROV has resulted in the sam-lite ontology[13] allowing the recording of specimen preparation chains via PROV [16]. As far as automatic data collection is concerned, internal processes are not always accessible for provenance recording, which is often the case with proprietary measurement devices hiding internal data transformations. In such cases, the specific information about the method and technology applied in measurement devices could be made available as contextual information via device type registries such as ESONET Yellow Pages[14]. Reliability of transmission is another aspect relevant for example in wireless sensor networks where collecting provenance becomes essential to ensure the integrity of the data packages transmitted [17].

Provenance is a crucial aspect in heterogeneous sensor infrastructures on the Web, also referred to as the Sensor Web, requiring the adaptation of existing data models. Integrating lineage information with observation descriptions may be done in a number of ways: Cox [16] for example aligned the Observations and Measurements model (O&M) with W3C PROV, while Jiang et al. [18] suggested directly extending PROV-O to cover O&M related concepts.

Increasingly, data are not acquired from one source but are derived from chains of services, resembling so-called "Virtual Data Products" (VDP). As an example for such cases, Yue et al.in [19] proposed the description of provenance via process models with a fixed structure which should be instantiated whenever a VDP is retrieved. This would enable prospective provenance based on the individual service descriptions and the process model or retrospective provenance derived from individual instantiations.

The tracking of provenance information during process execution can often be relatively straightforward, as many scientific workflow management platforms have already integrated this functionality based on provenance standards in their system. Example include Kepler [20], Pegasus [21], Taverna [22] (used by the LifeWatch[15] community) and dispel4py [23] (used in the seismology community). But if researchers run their processes on their private machines outside any particular provenance framework, then that provenance can only be tracked manually, which may be cumbersome and error-prone. One option is to use tools to extract provenance data from specially annotated scripts. Examples are the NoWorkflow system [24] for retrospective provenance and the

13 http://def.seegrid.csiro.au/static/ontology/om/sam-lite.html.
14 https://www.esonetyellowpages.com.
15 https://www.lifewatch.eu/.

YesWorkflow system [25] for prospective provenance which can easily be integrated into interactive notebooks like Jupyter/IPython [26] in use by many researchers today.

The use of the PROV standard by workflow management systems allows provenance information from multiple workflow management systems to be stitched together, but it is still challenging to produce a single cohesive provenance trace without some kind of overarching processing framework in place to orchestrate the provenance generation and storage. A promising approach is described in [27] where the Swift framework for parallel processing is augmented with provenance query frameworks such as MTCProv. Another possibility is to embed provenance recording at the operating system level as demonstrated by CamFlow, a Linux Security Module. Other approaches aim to wrap the entire process within a sandbox operating environment that enables the replication of the process via Docker virtual containers [28].

The PROV-AQ specification[16] provides recommendations related to the annotation of data objects with information on how to retrieve their provenance and to the discovery and query of PROV data. It expects that provenance is served via URIs provided via HTTP response headers, which either directly resolve to the provenance content or point to a dedicated query service. Another technology called Prov-pingback is a mechanism to track client derivations delivering URLs alongside each dataset which should be used to upload the provenance about the data transformations back to the provider [29].

The full visualization of provenance data as graphs of PROV-O triples may often not be satisfactory because of the potential complexity of data lineage. The provision of aggregated representations and thus large-scale overviews of the provenance information can instead substantially support users in the analysis of data generation.

Provenance Map Orbiter [30] is a technology that uses techniques for graph summarization, exploiting intrinsic hierarchies in the graphs, and semantic zoom. Other approaches are direct visualizations of subsets of the full provenance graph focusing on the temporal representation of chains of PROV activities linked together by the entities via Sankey diagrams [31]. Focusing on filesystem provenance, InProv [32] is a technology which transforms provenance graph data into temporally related aggregations visualizing them via a dedicated radial layout diagram. This approach allows navigation on the succession of different temporal visualizations including the storage of a visual protocol. It is being used in the seismology context in conjunction with the Bulk Dependency Visualiser [33] which provides large scale views on data dependencies in distributed stream processing environments such as data reuse between different users.

It is clear that there already exist a variety of tools for provenance gathering and visualisation, mostly based around a core group of standards that have been broadly accepted by the scientific community. It becomes necessary then to ask whether these tools and standards are seeing sufficient adoption in practice by the environmental and Earth science RI community, and if not, what the major barriers are to their adoption. In the following sections, we address how, in the context of ENVRIplus, we analysed the use cases and requirements of current European RI projects, and then drew upon the relevant standards and software libraries to provide a common provenance templating service for RIs and associated users.

[16] https://www.w3.org/TR/prov-aq/.

3 ENVRI RI Use Cases and Requirements

The provenance-related section of the questionnaire for the requirements elicitation process in the early phase of the ENVRIplus project (carried out in autumn 2015) intended to collect whether provenance was already considered in the data management plans of the RIs in the ENVRI cluster and if any related implementations were already in use [34]. Among the nine RIs that gave feedback, only two already had a data provenance tracking system integrated in their data processing workflows (EPOS and IS-ENES[17]). For those RIs where the latter was not the case, the next set of questions were focused on their potential interest in provenance: which type of information should be recorded, which standards to rely on and finally what sort of support was expected from ENVRIplus ICT staff. Most RIs considered provenance information as essential and some of them already stored provenance related information for certain aspects like data lineage following metadata standards such as ISO19139 or O&M. This information, however, was not considered sufficient to reproduce data since individual processing steps were not documented in enough detail. The outcomes suggested that it was highly relevant to learn more about what kind of information data provenance should provide, especially in contrast to what was already present in existing metadata about datasets. Another identified need was to get more insight into existing provenance recording systems.

As the outputs from the first requirement collection were rather moderate and unspecific, a second-round was undertaken in spring 2018 in order to retrieve more concrete information from the RIs (see Table 1). The objective was to understand the individual RI needs related to the potential implementation of provenance management systems. Regular teleconferences with live demos of implemented provenance services were thus offered to raise the awareness of the benefits and potential of data provenance techniques. Nine of 20 RIs sent their feedback, five of them have already participated in the first round of requirements collection, but this time giving a deeper insight into their needs, while the remaining four addressed this topic for the first time. As already anticipated, EPOS and IS-ENES, both quite advanced regarding this topic in comparison to other RIs, were able to provide more specific information about their requirements, but also about their existing implementations.

Table 1. Requirements collection (R1: 2015, R2: 2018).

	ACTRIS	AnaEE	EISCAT-3D	EMBRC	EMSO	EPOS	EURO-ARGO	EURO GOOS	IAGOS	ICOS	IS-ENES2	LTER	SeaData Net
R 1	x					x	x	x	x	x	x	x	x
R 2	x	x	x	x	x	x				x	x	x	

The RI representatives were asked to provide use cases with specific requirements considering that provenance information may be relevant in all phases of the research data life-cycle (DLC), from acquisition and curation to processing and use. Seven RIs provided specific use cases and requirements. Use cases were defined in this case as

[17] https://is.enes.org/.

descriptions of a set of interactions between a system and one or more actors, representing a user-perspective specification of functions in a system. For each use case, more than one requirement could be identified. The latter is understood as a functional perspective to approach the problem from a solution angle, providing a formal description of what users expect from the system [35].

The most evident differences in the provenance collection use cases and requirements between ENVRIplus RIs were found to be their varying focus in specific data life-cycle phases but also their varying level of automation. Some RIs included observation networks of scientists and/or instruments producing data (e.g. ACTRIS, EISCAT-3D and LTER-Europe), while others provided advanced processing services (e.g. AnaEE and IS-ENES2). Some RIs had fully automated sensor networks in place whereas human intervention was limited to monitoring, interpretation and/or maintenance tasks. Other RIs, in turn, encompassed considerably more manual steps occurring in the data acquisition but also during the processing phase. This diversity was clearly reflected in the use cases provided by the different communities.

In less automated settings, different aspects of provenance collection itself were reported and less on subsequent analysis and visualisation of such data. Respective use cases included scenarios for tracking lineage for script-based workflows, provenance for automated and non-automated data acquisition such as human observation and physical sample-based data collection, as well as provenance for data publishing and reuse.

In more automated settings (like in EPOS), the reported use cases were often addressing user needs and system features to address them, such as "discovery of experiments" or "navigation through data and dependencies" which were more relevant in the processing phase of the DLC.

Use cases mentioned by more than two RIs (highlighted in bold in Table 2) aimed at automated data collection via sensors, QC measures on instruments, data curation steps including QA/QC flagging procedures, data lineage of data products or aggregations as well as at model runs and their parameter settings.

As far as regards requirements, the different RIs converged more. Recurring requirements were various types of registries since recording provenance for processes with different agents and entities usually requires unique identifiers for each involved instance. Registries for any type of entity including persons, measurement sensors, software, etc. can thus be considered a prerequisite for any meaningful provenance approach. Other commonly expressed requirements were provenance tracking techniques, including domain-specific metadata from controlled vocabularies in the provenance tracks and recording of errata and of data use/citations [35, 42].

Based on the requirements of the various RIs in the ENVRI community and the resources available for development and innovation in the context of the ENVRIplus project, it was considered how best to support better provenance recording at a community level. With regard to this, a generic provenance service was developed that allowed for the generation of provenance traces based on pre-defined templates, which we will now describe in more detail.

Table 2. Use cases and requirements of RIs.

DLC phase	USE CASES	ACTRIS	AnaEE	EISCAT-3D	EMBRC	EPOS	ICOS	IS-ENES2	LTER
acquisition	method		x				x		x
	non-automated data collection						x		x
	non-automated physical samples						x		x
	automated data collection via sensors			x			x		x
	QC measures at instruments	x					x		x
	Eddy Covariance data algorithms						x		
	measurement station changes						x		
curation	**curation**		x	x		x		x	
	annotation		x						
	metadata		x						
	QA/QC	x		x			x	x	
	transfer to data centers							x	
	versioning							x	
publishing	**data products**				x	x	x	x	
	data lineage in scientific publications								x
	discovery of experiments					x			x
	interact. exploration of data dependencies					x			
processing	**model runs and configuration**		x	x		x	x	x	x
	data lineage in scripts			x					x
	track provenance in excel								x
	monitor workflow runs						x		
use	**data usage**				x	x	x		
	collaborative interactions					x			

REQUIREMENTS		ACTRIS	AnaEE	EISCAT-3D	EMBRC	EPOS	ICOS	IS-ENES2	LTER
provenance tracking			x	x	x	x	x	x	x
selective generation of traces						x			
ingestion of provenance								x	
errata tracking						x		x	
registries									
	datasets		x					x	x
	instruments/sensors	x				x	x		x
	physical samples						x		x
	persons					x	x		x
	sites/facilities		x						x
	lab equipment								x
	software/tools	x		x			x		
	publications						x		x
	vocabularies		x						
archives	long term data archival incl. provenance					x		x	

4 A Generic Provenance Service for the ENVRI Community (and Beyond)

As shown by the results of the ENVRIplus provenance use case gathering and requirements analysis, one main provenance-related distinction between research infrastructures can be drawn along the level of automation: highly automated research infrastructures, such as those operating on large-scale sensor networks, often feature dedicated software environments for executing clearly defined workflows, while less automated and smaller-scale infrastructures, such as those relying on human observation and sampling procedures, are often characterised by heterogeneous workflows consisting of alternating human and machine activities. The former is, therefore, better suited to becoming adapted for large scale provenance collection, while the latter represents a challenge in this regard.

Moreover, infrastructures are often still lacking important functionality required for meaningful provenance collection, an example being registries for relevant entities such as physical samples, sensors, instrumentation or personnel, important elements for the creation of provenance traces incorporating well defined and resolvable identifiers. Given the present scenario, the consideration of dedicated provenance services in the context of heterogeneous RI landscapes leaves the choice between concentrating on individual prototypes limited to a few selected research infrastructures only or on focusing on a more generic service concept suitable to a wide variety of RIs, ideally applicable to various levels of maturity. For ENVRIplus, the latter approach has been followed for being more in line with the overall project goals.

Generic approaches can, amongst other aspects, include means to create, store, query or visualise provenance collections. The latter three already require a collection of provenance data to be in place, suggesting that the creative aspect should be considered first when starting from scratch. Correspondingly, this also applied to the ENVRIplus context, motivating related activities accordingly. From an application-level perspective, there are three approaches to generating provenance [36]. "Passive Monitoring" refers to tracing a specific process solely based on the existing information it exchanges with its environment, not requiring any modifications of the original setup. "Overriding" is in turn about adding explicit provenance output to parts of the underlying execution environment (e.g. used software libraries) but not to the process itself, while "Instrumentation" refers to its direct provenance related modification. As far as the latter two are concerned, the heterogeneous landscape of environmental research infrastructures would thus require the direct modification of a wide variety of individual processes or underlying libraries, present either as compiled source code or via a scripting or workflow description language, for enabling the output of provenance. Although generic tools such as YesWorkflow [25] exist for annotating script-based code sequences, they do not cover the full range of possible workflow configurations and require deep knowledge of the code to be augmented. In turn, the notion of Passive Monitoring requires the identification of existing process output and its retrospective translation into a standardised form, potentially allowing the generation of provenance information without modifying underlying processes and their environments. Although having the disadvantage of being limited to the available existing output, this approach suggests itself as a low-threshold starting point for generating initial provenance traces for existing processes.

4.1 Using PROV-Template to Support the Generation of Provenance

One existing approach to turn existing process output into standardised provenance traces is called PROV-Template[18] [37]. As its name implies, it is based on the idea of creating templates which predefine the structure of the intended provenance information using variables which are later instantiated with appropriate data extracted from existing process output. As stated in [36], PROV-Template refers to prior descriptions of how retrospective provenance is to be collected and it is thus not related to the concept of prospective provenance outlined above. Closely related to the W3C PROV data model introduced in Sect. 2, the approach uses the model constructs specified there to define

[18] https://provenance.ecs.soton.ac.uk/prov-template/, retrieved March 6th 2019.

the templates, making them valid W3C PROV documents themselves. This, on the one hand, has the advantage that the outcomes of modelling activities to represent provenance traces for specific processes in PROV can be used as both templates and as a blueprint for implementing provenance output directly. On the other hand, existing libraries/services for storing, translating, manipulating, visualizing or validating PROV documents, such as the ProvToolbox[19] or the Python PROV library[20], can be applied to template documents as well.

PROV-Templates are instantiated via bindings which substitute template variables with actual values. The instantiation process is referred to as expansion, illustrated in Fig. 2 with an example template shown in the centre of the figure, representing an activity transforming one or more source datasets into a target dataset, featuring a responsible agent acting on behalf of an organization. As the "var:" prefixes of the element IDs suggest, they all serve as variables to be substituted with values extracted from the runtime log of a process corresponding to the template. This includes the mentioned PROV elements and their attributes, for which both keys and values can be specified as variables as well.

Example values for an appropriate process output are shown in the table at the top of Fig. 2 with the orange and green arrows indicating their bindings to the respective variables, resulting from a mapping effort which has to be done by suitable experts. The two different arrow colours emphasise that in this example, variables for attribute keys are substituted with column names and the remaining variables with column content, respectively. As visible in the table, the sets of substitute values for each variable can have different cardinality, such as for example in the columns "Source" and "Match Column" featuring the IDs of three source data files and the names of their data columns to be used for merging them, in which case the expansion results in multiple instantiations of the respective variables. While *n to 1* mappings such as the one presented in this example are expanded in a straightforward manner, the reader is referred to [37] for a formal description of the underlying expansion rules which also apply to more complex *n to m* mappings.

The result of the example instantiation is shown at the bottom of Fig. 2, illustrating how the expansion of the *n to 1* case yields three source file entities connected to the single transformation activity. Where no identifier is explicitly provided by the source data for a given element (such as for the activity variable itself), an expansion mechanism to provide an automatically generated unique ID can be used instead; this is indicated by the use of the "vargen" namespace. Although a useful feature for situations where the input data does not provide unique identifiers for certain elements, it has to be handled with care when considered for entities that potentially appear in multiple process instantiations, such as persons for example.

By design, PROV-Template enables the separation of concerns between the actual process and the generation of its provenance trace: as long as the process output contains sufficiently granular information in at least semi-structured form, it can be externally converted to W3C PROV via appropriate templates, relieving the process developers

[19] https://lucmoreau.github.io/ProvToolbox/, retrieved March 5th 2019.
[20] https://github.com/trungdong/prov, retrieved March 5th 2019.

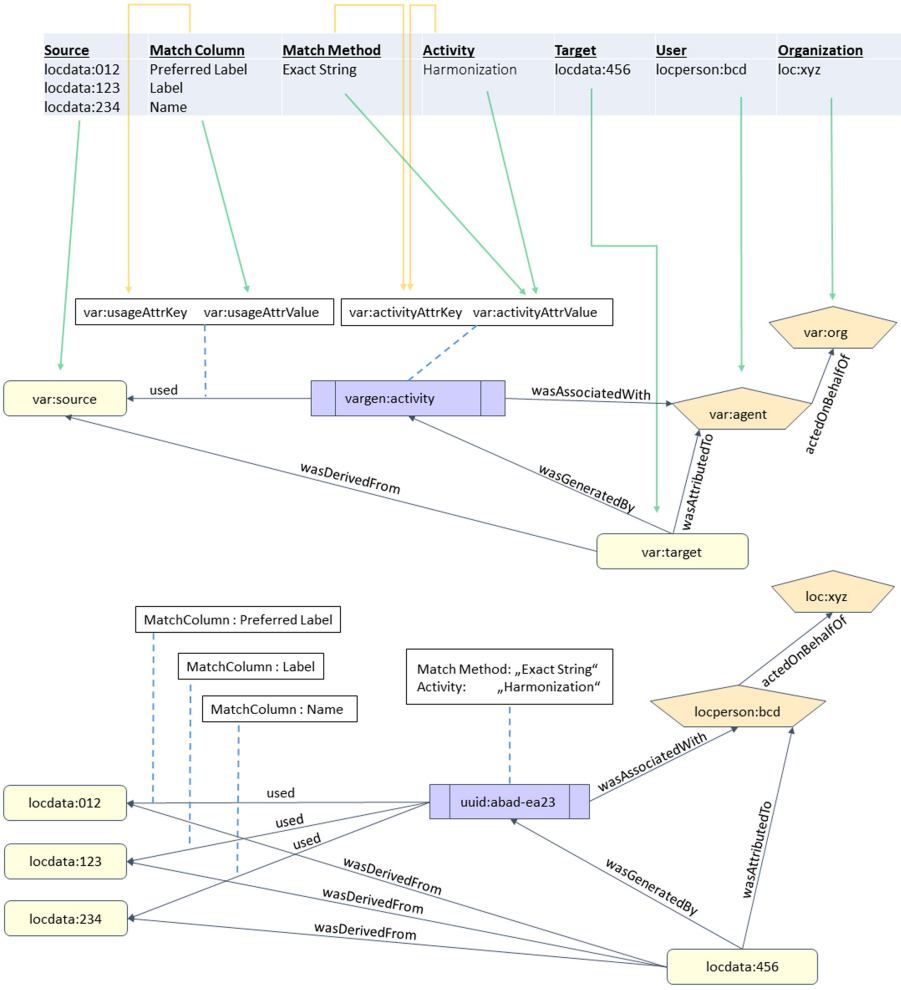

Fig. 2. PROV-Template expansion. (Color figure online)

themselves from the necessity to adhere to a specific data standard in this regard, potentially outsourcing the mapping effort to others. A resulting advantage is that intended changes in the provenance output at a later point in time can in many cases be achieved by modifying only the templates instead of having to touch the process implementation itself.

In the context of ENVRIplus, using PROV-Template appeared as a good starting point for provenance related activities. Assuming a general adherence to the W3C PROV standard, any modelling effort put into experimenting with templates wouldn't be lost even if other approaches than PROV-Template would be adopted in the end, since the created templates could then serve as a data model for any other endeavour to create PROV output. The suggested flexibility stemming from the separation of process output

and provenance creation led to the decision to consider the integration of PROV-Template into a community-wide provenance service. The resulting prototype consists of two main components: a catalogue for sharing templates and an attached service for their expansion. Their design is described in the following subsections.

4.2 A Catalogue for Environmental RI Related PROV-Templates

PROV-Template allows the design of a wide variety of patterns that can be applied to provenance related aspects of the full research data life cycle. Despite the identified heterogeneity of research infrastructures, it can be expected that many locally designed patterns are not only suitable for the specific infrastructure for which they were developed, but, with perhaps a small degree of customisation, also for other infrastructures with similar processes. In the context of ENVRIPlus, these considerations led to the development of an online service prototype dedicated to enable research infrastructures to upload, annotate and share their PROV-templates with the community. Templates shared that way should foster re-use and lead to more homogeneous and thus interoperable provenance representations.

Figure 3 shows the current version of the prototype[21] available online[22]. Its current content mainly serves as documentation for community experiments with the PROV-Template approach performed throughout the ENVRIplus project, with a more detailed description of these activities is available in [38]. The web interface consists of a scrollable list of uploaded templates, one per row. Each row features a rendering of the template as an SVG[23] graphic, allowing users to get a quick overview on its structure. Next to the rendering there is a dedicated section with descriptive metadata, currently consisting of basic Dublin Core[24] fields, and links to different W3C PROV serializations of the template. As visible at the top right of Fig. 3, users can log-in via existing social media accounts in order to upload new and manage existing templates. When registering a new template, users need to enter a minimum set of mandatory metadata fields and perform basic validation of the template data. A more thorough description of the steps required to upload and share templates is available as a dedicated manual [39].

Seen from a longer perspective, a catalogue service for PROV-Templates would benefit from various improvements. One important aspect would be the integration of vocabulary suitable for a more thorough description of templates in the context of their specific purpose within an RI's data life-cycle. It is expected that the use of more dedicated authoritative terminology would enable a more consistent annotation of templates, leading to better findability and the retrieval of more adequate templates for specific use cases. The ENVRI Reference Model described in [11] could serve as an important foundation in this regard and the integration of its fine-grained views on research infrastructures with the notion of W3C PROV-Templates potentially mutually beneficial. The availability of templates with fine-grained annotations would subsequently, however, require the adaptation of the search interface to efficiently make use of the increased

[21] https://github.com/EnvriPlus-PROV/ProvTemplateCatalog, retrieved March 7th 2019.
[22] https://www.envri.eu/provenancetemplates, retrieved March 7th 2019.
[23] https://www.w3.org/TR/SVG11/, retrieved March 7th 2019.
[24] http://dublincore.org/documents/dces/, retrieved March 7th 2019.

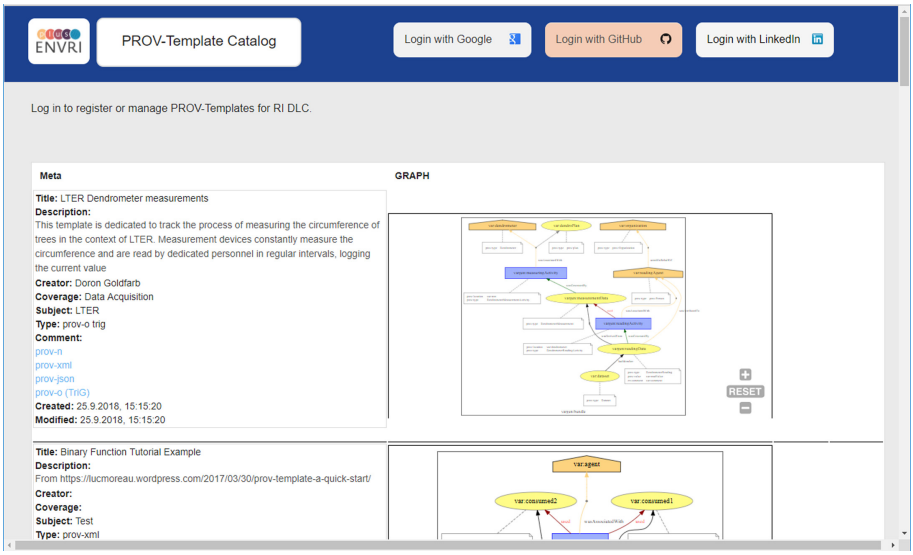

Fig. 3. ENVRIplus PROV-Template catalogue.

expressiveness. Another aspect for improvement would be community features such as rating, commenting and collaborative editing, potentially enabling users to go beyond mere re-use of each other's results.

4.3 Custom Expansion Service for PROV-Template

The second part of the ENVRIplus provenance service is dedicated to the expansion of PROV-Templates. Its basic component is a Python library[25] providing dedicated functions for translating a provided PROV-template and compatible bindings into an instantiated PROV document. Built on top of an existing Python library[26] for basic PROV handling, this implementation of the PROV-Template expansion mechanism is the first one of its kind and thus complements the Java-based proof-of-concept implementation available as part of the ProvToolbox.

The library follows the demand expressed by members of the ENVRI community for a way to directly integrate PROV-Template functionality into Python-based workflows without having to call an external service for that purpose. Besides being usable in standalone form, the library is nevertheless also integrated with the Template catalogue where it is encapsulated behind a dedicated web API, described in [39], for expanding the templates registered there.

5 Provenance and System Logs

A functional provenance service also requires other operations: provenance information capturing, storage, and query. Besides the provenance service presented in Sect. 4, we

[25] https://github.com/EnvriPlus-PROV/EnvriProvTemplates, retrieved March 7th 2019.
[26] https://github.com/trungdong/prov, retrieved March 8th 2019.

have also explored feasibility to link provenance information with the other types of information captured by the infrastructure and platform.

A complex scientific workflow often consists of multiple services, and those services are deployed on distributed infrastructures [40]. The runtime behaviour of the workflow, e.g. monitored by the underlying infrastructure, is important for analysing the workflow's provenance, in particular when the workflow has an unexpected performance issue or failure. However, the provenance and system metrics are provided by different information sources, which makes the integrated analysis difficult and time-consuming. It is thus challenging to analyse the workflow performance, due to difficulty in gathering and analysing performance metrics across distributed infrastructures.

A Cross-context Workflow Execution Analyser (CWEA) is developed for users to effectively investigate possible workflow execution anomalies or bottlenecks by combining provenance with available system metrics [41]. The tool is able to retrieve available system logs of the particular machines (virtual machines if in Cloud) and align them with the provenance provided by the workflow management system. In this way, a user (e.g. application developer or infrastructure operator) can inspect the infrastructure status for particular workflow execution, as shown in Fig. 4.

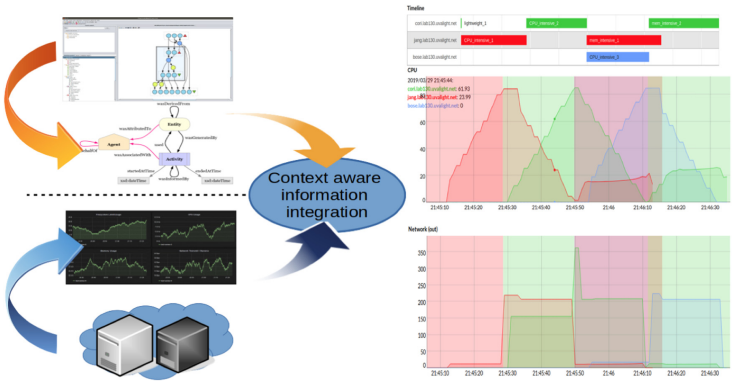

Fig. 4. The basic idea of the cross-context workflow execution analyser, and its output. In the right side of Fig. 4, user can interactively check the workflow processes (from the provenance), and check the system resource information (e.g. CPU and network).

6 Conclusion

In this chapter, we reviewed the state of the art of provenance tracking, focusing on provenance for research data and processes as needed for data-driven environmental science. The challenges of providing FAIR open data, particularly with regard to reproducibility, demonstrate a clear need for better and more extensive provenance gathering throughout the research data life-cycle. Much of the necessary research has already been accomplished, with the various methods, technology and standards ready to use in many

contexts and ready to roll out and adopt in others. There is still however a need for development to establish consistent implementations for every system, tool and context into which provenance must be situated. Some technical research into how to handle scale and security issues may be needed as this wider adoption occurs, as will the development of better governance frameworks and best practices for new researchers to adopt as part of their day-to-day activities.

In the context of the ENVRIplus project, a survey of provenance gathering capabilities and needs across the cluster of European environmental and Earth science research infrastructures was carried out. This provided the basis for the development of a shared provenance template service, via which RI developers and researchers can share executable specifications of the provenance patterns used within their infrastructures and workflows. This service also provided the ability to directly instantiate templates with uploaded datasets in order to automatically generate provenance traces in accordance with the W3C PROV standard. It is hoped that this kind of service can assist RI developers in formalising their provenance gathering procedures, share their work, and synchronise how provenance traces for a similar type of dataset and process are constructed across RIs, improving interoperability and reusability of the resources they provide to their respective scientific communities.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No 654182.

References

1. Moreau, L., Freire, J., Futrelle, J., McGrath, R.E., Myers, J., Paulson, P.: The open provenance model: an overview. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 323–326. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89965-5_31
2. Groth, P., Moreau, L.: PROV-overview. W3C. W3C Note, April 2013. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
3. Garijo, D., Gil, Y.: Augmenting PROV with plans in p-plan: scientific processes as linked data. In: CEUR Workshop Proceedings (2012)
4. Garijo, Y., Gil, G., Corcho, O.: Towards workflow ecosystems through semantic and standard representations. In: Proceedings of the 9th Workshop on Workflows in Support of Large-Scale Science, pp. 94–104. IEEE Press (2014)
5. Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttin, V., Ludäscher, B.: D-PROV: extending the PROV provenance model with workflow structure. In: 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP 13), Lombard, IL (2013)
6. Spinuso, A.: S-ProvFlow and DARE management for data-intensive platforms. In: RDA-Europe Meeting on Data Provenance Approaches, Barcelona, 15–16th January (2018)
7. Spinuso, A.: Active provenance for data-intensive research, Ph.D. thesis, School of Informatics, University of Edinburgh (2018)
8. Costa, F., et al.: Capturing and querying workflow runtime provenance with PROV: a practical approach. In: Proceedings of the Joint EDBT/ICDT 2013 Workshops, pp. 282–289. ACM (2013)

9. Bailo, D., Ulbricht, D., Nayembil, L., Trani, L., Spinuso, A., Jeffery, K.: Mapping solid earth data and research infrastructures to CERIF. Procedia Comput. Sci. **106**, 112–121 (2017)
10. Huynh, T.D., Moreau, L.: ProvStore: a public provenance repository. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 275–277. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16462-5_32
11. de la Hidalga, A.N., et al.: The ENVRI Reference Model (ENVRI RM) version 2.2 (2017). http://doi.org/10.5281/zenodo.1050349
12. Asuncion, H.U.: Automated data provenance capture in spreadsheets, with case studies. Future Gener. Comput. Syst. **29**(8), 2169–2181 (2013)
13. Zhang, Q., et al.: WIP: provenance support for interdisciplinary research on the North Creek Wetlands. In: IEEE 11th International Conference on e-Science (e-Science), pp. 521–528 (2015)
14. Buneman, P., Chapman, A., Cheney, J., Vansummeren, S.: A provenance model for manually curated data. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 162–170. Springer, Heidelberg (2006). https://doi.org/10.1007/11890850_17
15. Celino, I.: Human computation VGI provenance: semantic web-based representation and publishing. IEEE Trans. Geosci. Remote Sens. **51**(11), 5137–5144 (2013)
16. Cox, S.: Ontology for observations and sampling features, with alignments to existing models. Semant. Web **8**(3), 453–470 (2017)
17. Wang, C., Zheng, W., Bertino, E.: Provenance for wireless sensor networks: a survey. Data Sci. Eng. **1**(3), 189–200 (2016)
18. Jiang, J., Kuhn, W., Yue, P.: An interoperable approach for Sensor Web provenance. In: 2017 6th International Conference on Agro-Geoinformatics, pp. 1–6 (2017)
19. Yue, P., Gong, J., Di, L.: Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. Comput. Geosci. **36**(3), 270–281 (2010)
20. Altintas, I., Barney, O., Jaeger-Frank, E.: Provenance collection support in the Kepler scientific workflow system. In: Moreau, L., Foster, I. (eds.) IPAW 2006. LNCS, vol. 4145, pp. 118–132. Springer, Heidelberg (2006). https://doi.org/10.1007/11890850_14
21. Kim, J., Deelman, E., Gil, Y., Mehta, G., Ratnakar, V.: Provenance trails in the wings/pegasus system. Concurr. Comput.: Pract. Exp. **20**(5), 587–597 (2008)
22. Zhao, J., Goble, C., Stevens, R., Turi, D.: Mining Taverna's semantic web of provenance. Concurr. Comput.: Pract. Exp. **20**(5), 463–472 (2008)
23. Filgueira, R., Krause, A., Atkinson, M., Klampanos, I., Spinuso, A., Sanchez-Exposito, S.: dispel4py: an agile framework for data-intensive escience. In: 2015 IEEE 11th International Conference on e-Science (e-Science). IEEE, pp. 454–464 (2015)
24. Murta, L., Braganholo, V., Chirigati, F., Koop, D., Freire, J.: noWorkflow: capturing and analyzing provenance of scripts. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 71–83. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16462-5_6
25. McPhillips, T., et al.: YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. arXiv preprint arXiv:1502.02403 (2015)
26. Pimentel, J., Braganholo, V., Murta, L., Freire, J.: Collecting and analyzing provenance on interactive notebooks: when IPython meets noworkflow. In: Workshop on the Theory and Practice of Provenance (TaPP), Edinburgh, Scotland, pp. 155–167 (2015)
27. Gadelha, L., Wilde, M., Mattoso, M., Foster, I.: MTCProv: a practical provenance query framework for many-task scientific computing. Distrib. Parallel Databases **30**(5–6), 351–370 (2012)
28. Pasquier, T., et al.: Practical whole-system provenance capture. In: Proceedings of the Symposium on Cloud Computing, pp. 405–418. ACM (2017)
29. Lebo, T., West, P., McGuinness, D.L.: Walking into the future with PROV pingback: an application to OPeNDAP using prizms. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 31–43. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16462-5_3

30. Macko, P., Seltzer, M.: Provenance map orbiter: interactive exploration of large provenance graphs. TaPP **2011**, 1–6 (2011)
31. Hoekstra, R., Groth, P.: PROV-O-Viz-understanding the role of activities in provenance, in International Provenance and Annotation Workshop, pp. 215–220 (2014)
32. Borkin, M.A., et al.: Evaluation of filesystem provenance visualization tools. IEEE Trans. Visual Comput. Graph. **19**(12), 2476–2485 (2013)
33. Spinuso, A., Fligueira, R., Atkinson, M., Gemuend, A.: Visualisation methods for large provenance collections in data-intensive collaborative platforms. In: EGU General Assembly Conference Abstracts, vol. 18, pp. 14793 (2016)
34. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, Munich, Germany, pp. 551–556. IEEE (2015). https://doi.org/10.1109/eScience.2015.41
35. Magagna, B., et al.: Deliverable 8.5: data provenance and tracing for environmental sciences: system design, a document of ENVRIplus project (2018)
36. Frew, J., Metzger, D., Slaughter, P.: Automatic capture and reconstruction of computational provenance. Concurr. Comput.: Pract. Exp. **20**(5), 485–496 (2008)
37. Moreau, L.: A templating system to generate provenance. IEEE Trans. Softw. Eng. **44**(2), 103–121 (2017)
38. Goldfarb, D., et al.: Deliverable 8.6 Data provenance and tracing for environmental sciences: prototype and deployment, a document of ENVRIplus project (2018)
39. Goldfarb, D., Martin, P.: PROV-template registry and expansion service manual (2018). https://envriplus-provenance.test.fedcloud.eu/static/EnvriProvTemplateCatalog_Manual_v2.pdf
40. Zhao, Z., Belloum, A., Bubak, M.: Special section on workflow systems and applications in e-Science. Future Gener. Comput. Syst. **25**, 525–527 (2009). https://doi.org/10.1016/j.future.2008.10.011
41. el Khaldi Ahanach, E., Koulouzis, S., Zhao, Z.: Contextual linking between workflow provenance and system performance logs. In: 2019 15th International Conference on eScience (eScience), San Diego, CA, USA, pp. 634–635. IEEE (2019). https://doi.org/10.1109/eScience.2019.00093
42. Tanhua, T., et al.: Ocean FAIR data services. Front. Mar. Sci. **6**, 440 (2019). https://doi.org/10.3389/fmars.2019.00440