D8.5
Data provenance and tracing for environmental sciences: system design

WORK PACKAGE 8 – Data Curation and Cataloging

LEADING BENEFICIARY: Umweltbundesamt GmbH (Environment Agency Austria)

Author(s):	**Beneficiary/Institution**
Barbara Magagna	Umweltbundesamt GmbH (EAA)
Doron Goldfarb	Umweltbundesamt GmbH (EAA)
Paul Martin	UvA
Frank Toussaint, Stephan Kindermann	DKRZ
Malcolm Atkinson	University of Edinburgh
Keith Jeffery	NERC
Margareta Hellström	Lund University
Markus Fiebig	NILU
Abraham Nieva de la Hidalga	University of Cardiff
Alessandro Spinuso	KNMI

Accepted by: Keith Jeffery (WP 8 leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

A document of ENVRI^{plus} project - www.envri.eu/envriplus

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant

Deliverable due date: 30.04.2018/M36

Actual Date of Submission: 30.04.2018/M36

ABSTRACT

This deliverable reports the group efforts of Working Package 8 Task T8.3 on Inter RI data provenance and trace services during M24-36.

Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Markus Stocker	Technische Informationsbibliothek (TIB)
Robert Huber	UniHB

Document history:

Date	Version
16th April 2018	Draft for comments
26th April 2018	Corrected version
28th April 2018	Accepted by Keith Jeffery

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the author (Barbara Magagna barbara.magagna@umweltbundesamt.at)

TERMINOLOGY

A complete project glossary is provided online here: https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonize policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

D8.5 Data provenance and tracing for environmental sciences: system design

CONTENT OF DOCUMENT

What the reader of this report can expect is listed below:

- Definition and demarcation
- A description of relevant aspects of provenance management
- Challenges related to provenance
- A review of the latest technologies for data provenance
- An overview on existing standards and provenance models
- A description of international initiatives on provenance
- A collection of requirements and use cases of ENVRI Research Infrastructures
- Bringing this collection in relation to requirements and use cases of other initiatives
- A description of provenance practices of two ENVRI Research Infrastructures
- A description of ENVRI provenance related implementation cases
- How provenance relates to the ENVRI Reference Model and the OIL-E ontology
- How provenance relates to the ENVRI architecture
- A long term perspective on provenance

1 INTRODUCTION

The main objective of the T8.3 “Inter RI data provenance and trace services” is to provide an overview of common standards and best practices related to data provenance management and to increase the scientists’ awareness of the benefit of proper provenance information.

1.1 Motivation

Provenance has – over the last decade – gained some prominence in academic discussions surrounding curation, cataloguing and system architecture. There is a broad consensus that provenance is central to the requirement that scientific research should be reproducible. Yet only a few Research Infrastructures (RIs) have implemented provenance services or embedded tools in their system architecture.

Data (but not only, also software versions, workflows, etc,) are useful if accompanied by context on how they are captured, processed, analyzed, and validated and other relevant information that enables interpretation and use. This is what provenance is about. For users of data, the scientific basis of their analysis relies to a great extent on the credibility and trustworthiness of their input data. For publishers of data, the provision of provenance as part of their cited data is crucial for scholarship and reproducibility. A prerequisite to reproducibility of scientific conclusions is traceability. One must be able to trace from a conclusion in a publication back to the objects that were used to derive it.

1.2 Task description

The T8.3 “Inter RI data provenance and trace services” aims at introducing scientists and data architects to the topic of provenance. As there is no one-size-fits-all system for all domains and application areas, it investigates general provenance issues and provides the information needed to design provenance systems according to the explicit requirements of specific RIs. It helps defining specific needs and finding adequate answers to the challenges addressed. The task has to produce a report and a prototype demonstrator.

The deliverable at hand, D8.5 “System design”, is a report giving basic insights into relevant aspects of provenance management. It provides a thorough analysis of the state of the art and provenance models in use. It collects RI specific requirements and use cases and compares them to produce a synopsis of ENVRI provenance-related needs. This deliverable also provides general recommendations on short-term and long-term provenance.

This report takes common requirements of RIs throughout the entire data lifecycle into account: from acquisition, curation, to processing. It indicates standardised interfaces for querying, accessing and integrating provenance data and investigates arising provenance services for e-Infrastructure projects. Furthermore, it builds upon the semantic linking framework developed in T5.3 reusing existing standards, such as W3C’s PROV-O for possible general interoperability.

In the second phase of T8.3 the collected requirements and use cases will be further refined together with the RIs and referred to RDA provenance patterns, identifying state-of-the-art approaches on how to tackle the addressed issues. These activities will contribute to adoptions of provenance techniques and management practices for specific use cases to be integrated into the demonstrator.

1.3 Activities in the task

The activities for this task started with the requirements gathering in autumn 2015, which was organized around T5.1. A specific provenance related questionnaire was developed and distributed to the Research Infrastructures with the help of so called ICT-RI- go between to collect provenance specific requirements. Although the feedback was rather scarce, it made clear that there was a broad interest to get more insights into this topic. An initial state of the art analysis on provenance was undertaken and incorporated in D5.1 in March 2016.

The main activities started in autumn 2017 with a virtual Kick-Off meeting in October, which had the objective to prepare the Provenance Workshop held at the 6th ENVRIweek in Malaga in November. This workshop established a work plan with actions and contributions from RIs and ICT partners. Since then a template for requirements and use case gathering was developed, distributed to several RIs and the responses were collected during following months. Several teleconferences with provenance practices presented by EPOS/DARE and IS-ENES representatives followed. The 8.3 working team (ENVRI provenance group) participated in the regular teleconferences of RDA Provenance Patterns Working Group, the Workshop RDA-Europe Data Provenance Approaches held in Barcelona in January 2018, and the related sessions at the RDA Plenary meetings in Barcelona (April 2017) and Berlin (March 2018). At the site visits of EPOS in Rome in September 2017, of ICOS in Lund in December 2017 and of EISCAT/D4Science in Pisa in February 2018 ICT specific provenance requirements and implementations were presented and discussed. An implementation case was selected to demonstrate how provenance management can be applied. During the ENVRIweek in May 2018 a further selection of use cases for demonstration purposes and of existing provenance tools for testing purposes will be determined. Both selections will define the work framework for the second half of 2018 and lead to a demonstration on how provenance techniques can be deployed.

1.4 Layout

Because of the novelty of provenance services and the explicit wish of the RIs to get an introduction about provenance and related implementations before proceeding with the specification of their own individual requirements. In analogy to this, the deliverable starts with a state of the art analysis and review of provenance techniques before addressing the requirements of the RIs regarding this topic. The document is laid out as follows:

Section 2 – Background: Starts with a definition of provenance and with a review of technologies in this field. It is followed by a description of the W3C PROV standard and other provenance models and gives an overview of the ongoing initiatives advancing provenance technologies and implementations.

Section 3 – Requirements: Provides an overview of the requirements collected within the RIs and compares them with the collection of the RDA provenances patterns.

Section 4 – Best practices and implementations: Describes the ongoing developments related to provenance implementations within the ENVRI community.

Section 5 – Provenance related ENVRI Implementation Cases: Gives insight in the ongoing activities of ENVRIplus experts working on different aspects of provenance.

Section 6 - Recommendations: Provides guidelines for RI data architects and implementers how to implement provenance in their infrastructure and explains how this fits into the ENVRI architecture. Finally, the section gives an outlook of how provenance should be integrated in the long-term.

2 BACKGROUND

2.1 Definition and demarcation

Provenance, from the French term ‘provenir’ meaning ‘to come from’, was originally used to keep track of the chain of ownership of cultural artifacts, such as paintings and sculptures, as the chain determines the value of the artwork. Hence, the word provenance refers to the origins of an information bearing entity [Sweeney, 2008]. This seems to be the sole consensus on the different interpretations of the concept of provenance, which mainly depends on the context or domain. Thus, for archaeologists provenance is the place where an object was found whereas for geologists it is the reconstruction of the history of sediment movements over time.

2.1.1 Historic provenance principles

The origin of provenance goes back to the principle “Respect de fonds” [Duchein, 1983], also known as the “Principle of provenance”, established in 1841 by a commission of historians in France. After the French Revolution the need emerged to merge public and private collections of property, legal and historical records into a single national archive. For a period of fifty years archivists tried hard to re-arrange the documents according to a new imposed classification which resulted in chaos for the context of these records. The new principle prescribed instead that the archives of one creator had to be maintained separately from the archives of another creator and had to remain in the same original order (Strukturprinzip). Furthermore the additional contextual information collected by the original archivist was considered more important than the artifact itself. This relies on the assumption that the initial curator owns the most complete and accurate knowledge about the object of interest. A closely related concept is that of archival integrity, which holds that records emanating from the same source should be kept together. Although in the archival domain these conventions are discussed controversially because it implies that the fond is a stable entity whereas it is for sure subject to change these findings can be easily transferred to digital objects.

2.1.2 Definition

According to the definition of the PROV W3C Recommendation[1], a widely accepted standard and the backbone we are mainly referring to in this document, data **provenance is information about entities, activities, and people involved in producing a piece of data.** This information can be used to form assessments about its quality, reliability or trustworthiness. It is the entire information that ran through the whole history of data process, including all data sources and all processes generating these data. To be a source of trust it is essential that the evolutionary contexts are maintained according to the actual data lifecycle. To be reliable and of high quality the additional information about the digital object should be as comprehensive as possible and captured by the people directly involved in the data production. There should be one single provenance record (provenance summary) that integrates all fragments of provenance information tracked in the life span of the data production in each of the different distributed computational infrastructures where these processes have been performed.

[1] https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

2.1.3 Provenance and metadata

Provenance is often conflated with metadata. These terms are related, but they are not the same. Provenance records are a kind of metadata. There are other kinds of metadata that are not provenance. Descriptive metadata of a resource only becomes part of its provenance when one also specifies its relationship to deriving the resource. For example, a file can have a metadata property that states its size, which is an information that has nothing to do with provenance, since it does not relate to how the file was created. The same file can have metadata regarding its creation date, which would be considered provenance-relevant metadata. So even though a lot of metadata potentially has to do with provenance, the terms are not equivalent. In summary, provenance is often represented as metadata, but not all metadata is necessarily provenance [Incubator Group Report]. Thus provenance is a subset of metadata, which only contains information describing the lineage of data.

2.1.4 Provenance and logs

Provenance is not a new concept: activity logs have been maintained in many ICT systems in the past and the present. These activity logs vary in structure and detailed contents and there is no standard form. However, those associated with e.g. DBMS usually provide enough information for provenance purposes. Similarly some analytics / visualisation packages provide such tracking. In the GIS research area, for instance, the importance of process logs was recognized early on as they enabled users to reconstruct the processing steps that lead to the GIS result.

From lab notebooks used in chemistry and some aspects of bioscience, to ‘notebooks’ used in particle physics, to observation logs in astronomy the concept is well-understood. Increasingly analytical or observational equipment produce computerised activity logs. In general all the sciences have moved from handwritten to computer-based logs. The challenge lies in the reuse of log files for the reconstruction of data provenance. Tan [2017] developed a promising practice by modelling log events into pair-wise provenance relations using a multi-layered provenance model.

2.2 Basic aspects on provenance

Before addressing the state of the art in provenance theory and techniques we try to underscore the very basic elements that characterise provenance from the ICT perspective on RI needs.

2.2.1 Referable object

We need to be able to establish the artefact or object for which we want to collect provenance records. This may be any kind of physical, digital or conceptual thing. But to be able to refer to this thing it must be identifiable. This can be a web resource with an URI (Uniform Resource Identifier) or any object having a persistent unique identifier (PID). In this deliverable we are concerned about referable data objects only. These digital objects may be organized in collections, subgroups and thus provenance may refer to aspects or portions of an object. How the attribution to the PID should be related to a dataset and its granules is currently under investigation in various international working groups. There is yet no common agreement whether PIDs should resolve to some sort of semantic information (Research Object) [Bechhofer et al., 2010] or PID Information Types[2], to the dataset metadata or to a dedicated PID metadata

[2] https://rd-alliance.org/groups/pid-information-types-wg.html

kernel (DOI)[3]. The ability to describe the provenance of a dynamic, evolving resource is another challenge. Questions on how new versions of the resource should relate to one another and whether provenance records should be self-contained and attached to each incarnation or refer to prior ones for more details is discussed in chapter 5.1.

2.2.2 Uses of provenance

The underlying objective pursued in a provenance management system can optimize a specific usage of provenance. But once provenance records are consistently provided, at least the usages listed below can basically be enabled.

DATA REUSE

Reusability is one of the four foundational FAIR principles for scientific data management [Wilkinson et al., 2016], which marks the provision of provenance information as prerequisite. Bechhofer et al. [2010] distinguish different forms of reuse, such as reuse by sharing the whole data entity, reuse of constituent parts also called repurpose (such as methods referred to in the provenance chain), repeat the whole study in which the data were involved (may be hampered if the accessibility to the services required is restricted) or reproduce only the result by using provenance as replication recipe, whereas replay does not necessarily involve execution of services but enables the examination and understanding of the process.

DATA QUALITY

Lineage can help to estimate data quality and data reliability (trust) based on the source data and their transformations. It is also used for proof statements on data derivations. Trust may be also based on attribution information, by checking the reputation of the agents involved. Making the curation and processing steps evident by exposing provenance information increases data reliability and allows to assess the quality of the data. Users should be able to access and understand how trust assessments are derived from provenance.

ATTRIBUTION

This refers to the sources or entities that contributed to the creation of the data object. It can give credit and legal attribution to the people involved in data creation, enable the citation of data with the originator and determine liability in case of erroneous data.

INFORMATIONAL

A generic use of provenance is to query based on lineage metadata for data discovery. By browsing it, a context to interpret data is provided. Provenance-driven tools enable rapid exploration of results and provide insights into relationships between data which accelerates understanding.

COMPARISON

Scientists often have to compare the detailed process provenance and original entities of their experiments in order to better understand their results. Two results can be very different while their provenance indicates significant commonalities. Conversely, two outputs can seem alike and the provenance differs considerably.

DEBUGGING

[3] https://www.rd-alliance.org/group/pid-kernel-information-wg/wiki/pid-kernel-information-guiding-principles

Scientists can discover problems in the processes that generated the results by detecting failure symptoms in the provenance tracks.

AUDIT TRAIL

Provenance can trace the audit trail of data, determine resource usage and detect errors in data generation. The process that creates an audit trail runs typically in a privileged mode, so it can access and supervise all actions from all users. This makes not only the data lineage transparent but also the use of data after their publication, which could expose sensitive and personal information. This might require some form of anonymisation for privacy or identity protection reasons. It is questionable if usage tracking should be a by-product of provenance which normally should just focus on the origins and transformations of the data product rather than on its users [Bier 2013].

2.2.3 Provenance life cycle

Ideally, all actions performed on data during their lifecycle are captured via provenance services. But where does the life cycle start? According to Fleischer & Jannaschk [2011] provenance capturing should already start before a digital object is produced; it should start at the point of origin – at the research sites including all human activities and analytical procedures involved in the observation of a natural phenomena or object.

Similarly to the lifecycle of data, also provenance metadata has its own phases with its own requirements, techniques and practices. We distinguish between *provenance capture* (analogue to data acquisition phase), *provenance storage* (analogue to data curation), *provenance access and query* (analogue to data publishing), *provenance analytics and visualization* (analogue to data processing and data use). While data and their provenance metadata should be made available at the time of data publication, provenance capture could be extended to all phases of the data life cycle to track each step of the entire evolution including the generation of data, and data transformation during the curation phase. After the publishing phase capturing provenance metadata on data processing and data use can be of important value. This requires approaches supporting heterogeneous workflow systems that map distributed provenance archives into a centralised record to enable provenance summaries of a digital object based on a standard model such as W3C PROV.

2.2.4 Types of provenance information

This chapter provides an overview of terms used in literature[4] and often referred to in this document in order to specify the type of provenance meant.

RETROSPECTIVE PROVENANCE

Also known as r-prov, retrospective provenance refers to the provenance of data and captures the steps that were executed in the generation of a data entity, recordings of the base inputs and intermediate data entities involved, and information about the associated environment. This information can be recorded at varying levels of detail and granularity.

PROSPECTIVE PROVENANCE

Also known as p-prov, prospective provenance describes the abstract representation of the procedure or "recipe" required to generate a data entity. It corresponds to the workflow

[4] http://vcvcomputing.com/provone/provone.html

specification, which details the steps involved in the process. It increases the understanding for the process and secures the quality of the data produced in accordance with the plan.

PROCESS PROVENANCE

Process provenance represents the evolution of the workflow description. It helps understanding the various changes that were carried out to generate the desired data products.

PROVENANCE OF DATA STRUCTURE

This type of provenance information refers to the most relevant aspects of how the data used and/or produced by a workflow is organized. For scientific workflows this includes the inputs and outputs of the various tasks within the workflow.

2.2.5 Challenges

EFFICIENCY OF PROVENANCE COLLECTIONS

Collecting provenance information in data-intensive environments in a way that supports complex queries can be challenging. The data ingestion phase in streaming workflows requires the raw data files to be sliced into many small and volatile data chunks, which often are unpreserved and thus not reusable. This may cause an explosion in the number of provenance traces produced. Nevertheless provenance data should be accessible at runtime, during harvesting, to enable intermediate interaction by users [Spinuso 2018].

GRANULARITY

The scale of provenance information is a critical issue, as the size of the provenance records may exceed the scale of the artifacts themselves. Tradeoffs between the granularity of the provenance information kept and the actual amount of detail needed by users of provenance must be made. The granularity at which provenance is captured determines the cost of collecting and storing the related contextual information. The range spans from provenance on attributes and tuples in a database (fine-grained provenance) to provenance of collections of files (coarse-grained provenance). However, both storage of fine-grained provenance can be managed to be smaller if it is compressed.

REPRESENTATION OF DOMAIN SEMANTICS

The data analytical power of provenance depends on the expressiveness of domain-specific data along the provenance traces. It may be difficult to reference domain-specific data in a query if individual domain-data values are not represented in coarse-grain provenance. The representation of domain semantics is broadly discussed in [Sahoo 2011] and [De Oliviera et al. 2015] and touched upon in chapters 2.4.3 and 2.4.7. According to Spinuso [2018], processes and data must be described by precise metadata including scientists' annotations, which are a flexible way of describing application and domain related features. However, both domain semantics and annotations are not addressed in standard provenance models such as W3C PROV[5].

INTEROPERABILITY

Efficiently storing and ensuring seamless propagation of provenance may be challenging as data are often transferred across heterogeneous systems and different representations or area used across multiple applications during their lifetime.

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

INCOMPLETE AND UNCERTAIN PROVENANCE

Provenance information is generated in many different ways often in multiple systems or alternatively just be tracked in specific phases of the data lifecycle (e.g. during processing). This may lead to fragmented and incomplete provenance. It is thus necessary to find ways to produce integrated summary provenance out of such fragments [Missier, 2016]. Provenance may also be provided with some uncertainty or be of a probabilistic nature caused by limited access to the complete provenance information or by problems occuring during the provenance tracking process.

TRUST

Provenance files are considered a form of evidence for the associated data. Thus, it is necessary to ensure that the provenance itself can be trusted, e.g. not to have been tampered with. Blockchain is one of the technologies that could address this issue as outlined in Stoffers [2017] (see also chapter 2.4.11).

EASE OF USE

To achieve understandability and usability it is important to provide means to allow for multiple levels for abstraction in the provenance records of a data object, well-arranged presentation and visualization of provenance information other than just a set of provenance statements. Facilitating the findability of provenance information and query formulation and executions by appropriate services and tools at hand becomes increasingly important with the availability of provenance information.

2.3 Technology review

A growing number of technologies exist for for dealing with different aspects of provenance, from collection and storage to visualisation and use. Many of these technologies exist as extensions or outgrowths of prior technologies for databases or for workflow management that have a clear need for well-ordered provenance recording. Increasingly however there are now also technologies being developed for the handling of provenance data irrespective (ironically) of their provenance, based around mature standards that provide a common way of describing such data across many different use cases and domains.

In this section, we review some of the technologies that are now available to developers, infrastructure engineers and researchers for producing, managing and making use of provenance records. We base our approach broadly on the research data lifecycle espoused by the ENVRI Reference Model, starting from acquisition through curation, publishing and processing up to visualisation and use. The respective data processes in environmental research span across data acquisition via measurements, observations and samples to data generation and processing via simulation and modelling in scientific workflows. The collection/creation of related provenance data should start as early as possible within this process. While the discussion of individual provenance related issues is out of scope here, it is nevertheless important to distinguish between automated settings, i.e. sensor based data collection or scientific workflow based data modeling and simulation, and manual scenarios, i.e. data acquisition via human observation or lab based processing of samples. Each of the two scenarios comes with specific provenance related issues which must be taken into consideration.

2.3.1 Provenance for manual processes

A main scenario for provenance in manual processes is data collection, which is performed either using traditional “analog” means such as pen and paper, or directly via digital devices such as handheld computers, etc. Using the former requires an additional step where the “analog” information is transferred to its digital counterpart, such as a spreadsheet or database, thus introducing an additional source for errors into the process.

As far as data entry in Spreadsheets is concerned, [Asuncion, 2013] provided an add-in called “InSituTrac” for the collection of data provenance from within Excel. Integrated into the Excel UI, the tool enables users to record their actions into a provenance log, which can later be queried and visualized. Two case studies supported the claim for the usefulness of this approach.

Since Spreadsheet integrated provenance collection does not eliminate the potential need to start tracing right at the spot where observations are recorded, approaches such as EcoProv, discussed in [Zhang et al., 2015], went one step further and aimed to completely replace pen-and-paper approaches with handheld digital devices, requiring users to log-in and thus allowing the recording of individualized provenance information ”right away”. The authors argued that in the future information such as geographic location and current weather conditions, etc. could be directly derived from the handheld device’s location context and automatically entered alongside the observation data. Since observation based data collection often takes place in remote locations without Internet access, the ability to enter information while being offline nevertheless remains a strict requirement.

In general, manual data collection and the subsequent entry into Spreadsheets/database forms is an error prone process whose confidence could be greatly increased via the recording and provision of relevant provenance information. As with all processes at the human-computer interface level, however, the success greatly depends on the willingness of the users to contribute. A general necessity is thus to design the respective interfaces to be as unobtrusive as possible, hiding away the actual provenance collection in the maximum possible way.

MANUALLY CURATED SCIENTIFIC DATABASES

As stated in [Buneman, 2006], manually curated scientific databases increasingly replace printed sources for raw scientific data, especially in the field of Bioinformatics but also in other domains. Since such data collections are usually composed of contributions from many different sources, their tracking is essential for maintaining scientific integrity. Addressing this development, the authors provided the so-called ‘copy-paste model” designed to capture chains of insertions, deletions and importations from sources to a target database. A proof-of-concept implementation was provided and the authors stated that the integration of their approach into typical Web-based workflows, such as manually browsing Web representations of source databases and copying relevant content into a respective Web based target database entry form, could be realized via the conversion of Web browser activity logs into dedicated provenance records.

USER GENERATED CONTENT/” HUMAN COMPUTATION”

A recent trend in data generation which is relevant in the context of provenance is User Generated Content (UGC) and the notion of ‘Human Computation”, often referred to as ‘Citizen Science” in a scientific context. Crowd knowledge is increasingly used in various application scenarios, either voluntary, such as in Wikipedia or Open Street Map, via ‘Games with a purpose” (GWAP) or even via dedicated marketplaces such as Amazon’s Mechanical Turk. Given the recent interest in and emerging success stories about introducing Citizen Science into Environmental Science scenarios [McKinley et al., 2017], it will be necessary to consider provenance related issues in this regard in the future.

The overall process of involving crowd-workers for specific tasks is the formulation of the task, its execution via the crowd and the subsequent aggregation of the results based on dedicated mechanisms such as majority voting. Aggregation is necessary since in most cases, one and the same sub-task is assigned to more than one individual to gather a second (or third, etc.) ‘opinion” in order to rule out errors. The aggregation step compares the results achieved by different individuals for the same sub-task and decides which result should be used. Provenance can be used to trace the genesis of UGC back to the individual contributions, thus improving data integrity, but also enables the creation of trust indices for individual crowdworkers, e.g. by counting how often their results were favored during previous aggregation steps.

[Celino, 2013] described such an approach in a Voluntary Geographic Information (VGI) setting, where a GWAP called ‘Urbanopoly” was used to playfully collect information about geographical features. Using a PROV-O extension as underlying model, the authors traced each user’s contributions and ranked them via a score function based on necessary effort, contributor reputation (derived from previous activity tracked via PROV) and geographical distance to the feature.

COMBINATIONS OF MANUAL AND AUTOMATIC DATA PROCESSES

One case where manual and automatic processes often coincide is sample-based data collection and processing. River water samples are for example usually taken and preprocessed by human personnel, while the subsequent analysis takes place in laboratories using automated equipment. Provenance tracking should ideally start with the human sampling process and subsequently continue with more automated workflows. This requires a consistent infrastructure for providing sample identifiers as well as a registry for used equipment in the more automated steps as outlined above. As far as data structures and models are concerned, [Cox, 2017] for example described the alignment of the ISO 19156 Sampling Features Schema with W3C PROV in form of the sam-lite ontology[6], allowing the recording of specimen preparation chains via PROV.

2.3.2 Provenance for automated processes

While automated data collection and processing suggests that the recording of related provenance information is relatively straightforward compared to the manual cases, specific challenges must nevertheless be taken into account. This is especially relevant with respect to the transparency of processing steps taking place internally in measurement and analysis hardware/devices, e.g. converting raw measurements to data values via aggregation, but also issues regarding data integrity during transport must be considered. Scientific Workflow management systems on the other hand often offer internal means to collect provenance data during execution, but in many cases research processes take place in rather ad-hoc setting outside such systems, e.g. in form of manually “stitched” combinations/chainings of various scientific scripts. This subsection discusses some important aspects in this regard.

PROPRIETARY MEASUREMENT DEVICES

The internal processes within proprietary measurement devices are often not accessible and a complete trace of all internal data transformation steps thus not possible. In such cases, related

[6] http://def.seegrid.csiro.au/static/ontology/om/sam-lite.html

provenance information would be limited to the type and serial number of the device, ideally including information about its internal configuration and data processing, as well as information about recent calibrations, etc.

While the provision of maintenance metadata such as calibration protocols etc. mainly lie in the responsibility of the Site maintainer, device specific information which can be used for annotating original data about the method and technology of its own genesis could be made available via device type registries, such as for example provided via the ESONET Yellow Pages[7] for Deep-Sea-Observatories.

LOW LEVEL PROVENANCE EXAMPLE: WIRELESS SENSOR NETWORKS

Provenance can be collected at very different levels, ranging from very low-level approaches capturing the underlying processes at hardware, operating system and below OSI session layer level, to high level descriptions of the actual processes built on top. As an example for a low level provenance collection scenario, Wireless Sensor Networks (WSN) are increasingly considered an alternative to Satellite based remote sensing or Cell phone infrastructure based data networks [Corke et al., 2016].

As reported in [Wang et al., 2016], Wireless Sensor Networks can easily be tampered with and collecting provenance is thus essential for ensuring integrity on data packet level. Since wireless networks often require intermediate nodes for long distance data transmission, however, the amount of collected provenance data can easily exceed the volume of the actual measurements. The authors discuss a number of techniques to address this issue in different network scenarios.

WEB SERVICES/SENSOR WEB

Approaches such as the one outlined in [Wang et al., 2016] deal with packet (low-) level provenance and are thus mainly concerned with data integrity during transmission. These views, however, are usually hidden away from regular users. Higher levels require a different type of provenance, describing the used sensors and the respective research scenarios, conditions, involved agents, etc. in order to track the genesis of the data. Especially in recent years, scientific data made available via Web services are increasingly sourced into scientific workflows, making it necessary to provide means to track composite provenance.

In the context of data acquisition, the concept of the Sensor Web has emerged as approach to provide an interoperable, Web service based layer on top of heterogeneous sensor infrastructures. This is supported by recent standards such as the OGC Sensor Observation Service (SOS)[8] which is itself built on dedicated models to describe sensors (SensorML[9]) and observations/measurements (Observations & Measurements[10]). Provenance is a key aspect in this regard and has already been included in the respective underlying ontologies via alignments to W3C PROV, seeking to allow a seamless integration of lineage information with the description of observations [Cox, 2017]. Other approaches such as [Liang et al., 2017] in turn suggested to extend PROV-O towards embracing concepts from vocabularies such as O&M, seeking to allow provenance queries in order to answer observation related questions (Why, How, Who, Where, When, What). The authors mentioned different provenance views, e.g. considering a sensor as an agent with respect to the observation process but as an entity with

[7] https://www.esonetyellowpages.com, retrieved March 20th, 2018
[8] http://www.opengeospatial.org/standards/sos
[9] http://www.opengeospatial.org/standards/sensorml
[10] http://www.opengeospatial.org/standards/om

respect to its own deployment. They also provided SPARQL query examples in this regard, asking for the provenance of specific observations as well as the identification of all sensors deployed after a specific date. As the example for the different views on sensor provenance suggests, it will thus require rich temporally-bound semantics to distinguish between such aspects.

In more general cases, external data are not derived from one single (type of) source but resemble 'virtual data products" (VDP) composed of contributions of chains of services. Such service chains can be described via process models instantiated whenever a VDP is retrieved. The served content behind the VDP has a fixed structure, but the actual values can change (e.g. 'last month's temperature curves"). It is thus possible to provide prospective provenance derived from the individual service descriptions and the process model, or retrospective provenance derived from individual instantiations. Since data often have to be transformed in order to be useful for a specific consumer, transformation routines have to be identified, applied and logged as provenance. Prospective provenance describing the service chain behind a VDP can for example be used to identify such necessary data transformations for specific usage scenarios. [Yue et al., 2010] described such a system in the context of geospatial Web services, deriving provenance information from chains of GIS Web services via an approach the authors referred to as metadata tracking. The proposed system made use of Semantic Web technologies for the service descriptions as well as the representation, creation and storage of the resulting provenance information.

SCIENTIFIC WORKFLOW MANAGEMENT AND EXECUTION

The difficulty of generating provenance during process execution (describing the process, the processing environment, the effect on existing data or the link to new datasets, etc.) is largely dependent on whether or not provenance tracking is already integrated into the processing platform, and in what form that tracking takes place. For example, if researchers operate within an integrated virtual research environment using some established scientific workflow management system to coordinate and execute their data analytics and other processing tasks, and if that workflow management system already contains a provenance subsystem, then the extraction and recording of properly-structured provenance data is (in principle) very simple. On the other hand, if researchers are running processes on an ad-hoc basis on their own private machines outside of any provenance framework, then it befalls upon them to reconstruct the provenance record manually before uploading their results to a repository (assuming they even go that far)—this is much more difficult and error-prone.

Many scientific workflow management systems support provenance, including (but in no way limited to) Kepler [Altintas et al., 2006], Pegasus [Kim et al., 2008], Taverna [Zhao et al., 2008] and dispel4py [Filgueira et al., 2015] (used in the seismology community). The provenance capabilities of these systems are typically focused on being able to perform 'smart' re-runs of workflows, but the outputs are typically in accordance with provenance standards, allowing for a broad range of analyses via various tools, for example to characterise activity on the underlying e-infrastructure [Madougou et al., 2013], often agnostic of the original workflow management system, such as in [Costa et al., 2013]. The use of a workflow management system confers a particular benefit for distributed computing; it is difficult to produce a cohesive provenance trace for a process workflow involving multiple machines and parallel execution without some kind of overarching processing framework in place to orchestrate/choreograph the provenance generation and collection. Parallel processing frameworks as Swift can be augmented with the use of provenance systems such as MTCProv [Gadelha et al., 2012], a provenance query framework for many-task scientific computing. Likewise, the Century stream processing infrastructure for supporting online healthcare analytics [Misra et al., 2008] uses Time-Value-Centric (TVC) provenance [Wang et al., 2007]. Given the onerous requirement to store all data streams for TVC provenance, [Misra et al., 2008] proposed a new hybrid provenance architecture, called Composite Modeling with Intermediate Replay (CMIR) that uses data replay to recreate data elements of streams internal to the processing elements in the stream processing workflow.

A notable benefit of the use of common provenance standards by workflow management systems is the allowance for ‘stitching together’ of provenance traces from multiple workflow management systems (e.g. Kepler and Taverna [Missier et al., 2010]) in order to gain a more holistic view of computational science workflows, which can then be subject to fine-grained query of workflow provenance traces [Missier et al., 2008].

Fortunately, the use of workflow systems is already established within the environmental sciences, supported by many research infrastructures. LifeWatch makes use of Taverna, for example, while the VERCE project (operating as a contributor to EPOS) implemented provenance facilities based on W3C PROV linked to the dispel4py workflow description framework [Atkinson et al., 2015], which can be queried via a custom provenance explorer GUI as described below. Nevertheless, this is not sufficient to cover all needs of environmental scientists.

As already discussed, many scientists do not operate within the confines of a particular workflow system or data processing platform, preferring to run their own scripts, typically in their own environment (e.g. their office laptop). In this case there are still ways to (partially) automate the generation of provenance data. One way is to use tools that extract provenance data from specially annotated scripts, e.g. the NoWorkflow system by [Murta et al., 2015] for retrospective provenance and the accompanying YesWorkflow system by [McPhillips et al., 2015] for prospective provenance. Tools like NoWorkflow can easily be integrated with interactive notebooks like Jupyter/IPython [Pimentel et al., 2015] to better explore/visualise the collected provenance data, without requiring that scientists import their code into a particular workflow management system. As part of ProvToolbox, tools such as PROV-TEMPLATE take a declarative approach to defining provenance templates for integration into code to ensure the recording of provenance according to the PROV standard [Moreau et al., 2018]. The role of such tools are to make the adoption of formal provenance tracking much simpler for researchers and developers.

If researchers are not using code-level annotation to facilitate provenance gathering, nor are they working within a processing framework that has an integrated provenance tracking system, then another possibility is to instil provenance recording at the operating system level. As an example, CamFlow[11] attempts to embed provenance capture into the operating system, in this case Linux as a Linux Security Module [Pasquier et al., 2017]. Conversely, it should be remembered that one of the major justifications for provenance collection is the reproducibility of processes. While most provenance research focuses on creating ‘traces’ of provenance based on collecting and characterising (via metadata) intermediary steps and results in the process workflow, other approaches include wrapping the entire process within a sandbox operating environment that itself can be packaged and re-executed to replicate the process when needed, e.g. using Docker virtual containers [Meng et al., 2015].

[11] http://camflow.org/

2.3.3 Provenance data management, discovery and retrieval

There exist a number of core standards for provenance, though the two highest profile standards are the Open Provenance Model (OPM) [Moreau et al., 2008] and W3C's PROV recommendation. The W3C PROV Recommendation [Groth and Moreau, 2013] consists of a number of constituent standards including for PROV XML and for PROV as an ontology for RDF-based data (PROV-O). Other formats for PROV data have been proposed including in JSON-LD [Huynh et al., 2016]. Aside from the direct description of provenance traces, specific provenance models may require specific query languages in order to better extract useful information from those provenance traces. An example of an OPM-based query language is OPQL [Lim et al., 2011a], while an example of an OPM-based relational database oriented scientific workflow provenance system is OPM-PROV [Lim et al., 2011b]. One advantage of PROV (specifically the use of PROV-O) is the ability to query PROV traces using SPARQL and other RDF data query frameworks, though even there a higher-level query language that provides an abstraction layer over the underlying SPARQL query (particularly for more complex joins of data that might be needed) would still be helpful for researchers wishing to engage with provenance data. Similarly, standard triple store databases can be used to store PROV-O data in a reasonably natural way.

A number of useful tools have been made available online for use by researchers and other users of provenance data in order work to with these standards. Examples can be found on sites such as https://openprovenance.org/ and https://provenance.ecs.soton.ac.uk/, which provide public provenance data storage based on ProvStore[12] [Huynh and Moreau, 2014], as well as validation against the PROV standard and conversion services for various standard output formats. ProvToolbox[13] is a Java library for manipulating provenance descriptions (meeting the PROV standard), and converting them between RDF, PROV-XML, PROV-N, and PROV-JSON encodings. Similar libraries have been developed for other languages such as Python[14] and R[15].

Many open scientific data repository initiatives are taking provenance into account, either explicitly following the PROV standard or otherwise codifying their provenance traces via libraries such as recordr[16]. Such initiatives include DataONE [https://www.dataone.org/] and the Dataverse project[17]. Many make use of the notion of 'research objects' to describe research activity and datasets, which provides a natural context to attach provenance data based on a standard ontology [Belhajjame et al., 2015] - such an approach can be taken to attach e.g. PROV-O documents to a contextualised data package.

DISCOVERY & RETRIEVAL

In order to be readily available for analysis or accumulation with other sources' lineages, provenance data must be provided in a findable and accessible way. Dedicated services to store provenance documents, such as ProvStore already described above, often include individual means to discover and view hosted resources, such as via dedicated REST API. Specifications regarding discovery and access can also be found alongside existing provenance standards. In the context of W3C PROV for example, the PROV-AQ specification[18] describes simple ways to

[12] https://provenance.ecs.soton.ac.uk/store/
[13] https://lucmoreau.github.io/ProvToolbox/
[14] https://pypi.python.org/pypi/prov
[15] https://github.com/End-to-end-provenance/RDataTracker
[16] https://www.rdocumentation.org/packages/recordr/
[17] https://dataverse.org/
[18] http://www.w3.org/TR/prov-aq/

annotate data objects with information how to retrieve their provenance and also provides recommendations regarding how to discover and query PROV data in more elaborate cases. The approach basically assumes that provenance is always served via URI, which can either directly resolve to the provenance content or point to a dedicated query service. For either mode, URIs shall be provided via HTTP response headers and/or embedded in HTML or RDF content. SPARQL queries to the underlying provenance served via Triplestore represent the most elaborate way how to interact with PROV data according to PROV-AQ, although the specification does not explicitly require queries to have to be imperatively based on PROV-O vocabulary terms.

Another aspect of provenance discovery and retrieval is the tracking of client derivations. In order to provide data hosters with mechanisms to foster the reporting of what has been done with their data, the W3C PROV specification features a mechanism called Prov-pingback[19]. Its basic concept is to deliver a specific URL alongside each provided dataset which clients shall use to upload the provenance about their own data transformations back to the provider. This enables institutions to maintain a central service for discovering a dataset's "offspring". An example for the implementation of such a service is described in [Lebo et al., 2014].

2.3.4 Provenance visualization

The (interactive) visualization of generated provenance data supports users in analyzing different steps and other aspects in the traced scientific processes, allowing them to get a large scale overview potentially providing insights on macro level which would be impossible to gain otherwise.

Assuming the basic structure of contemporary provenance data to be in form of a graph (e.g. via PROV-O conforming triples), the main distinction between existing approaches can be made by whether the graph data are visualized directly or transformed into a different, e.g. aggregated, representation first.

DIRECT VISUALIZATION OF PROVENANCE GRAPH

Due to the potential complexity of large provenance graphs, their direct visualization can easily exceed the capacity of the medium (paper, screen) and/or the viewer. A number of different techniques have therefore been applied in order to tackle this visual overload.

[Macko and Seltzer, 2011] described an approach called Provenance Map Orbiter, applying two interrelated techniques the authors referred to as graph summarization and semantic zoom. Summarization was achieved by harnessing intrinsic hierarchies usually present in provenance graphs such as e.g. process invocation trees in order to collapse graph nodes based on their hierarchical level. The zoom feature in turn enabled users to start with a highly summarized representation of the provenance graph and to drill down into the hierarchical details via simple zooming, shown in Figure 1. This way, the authors claimed to enable the exploration of graphs having up to 10^5 nodes. Larger graphs could be reduced via filtering of node attributes, lineage queries such as selecting all descendants of a specific node were supported as well.

[19] http://www.w3.org/TR/2013/NOTE-prov-aq-20130430/#provenance-pingback

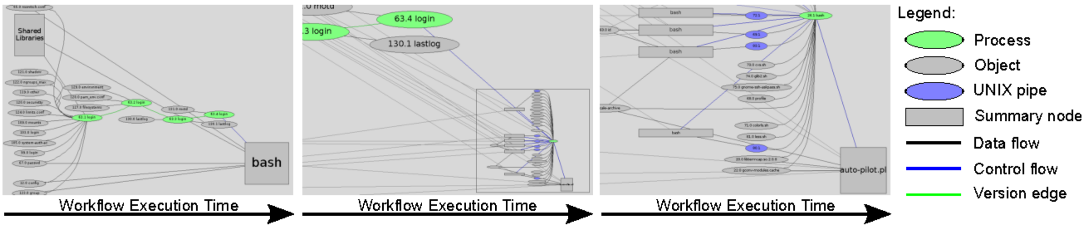
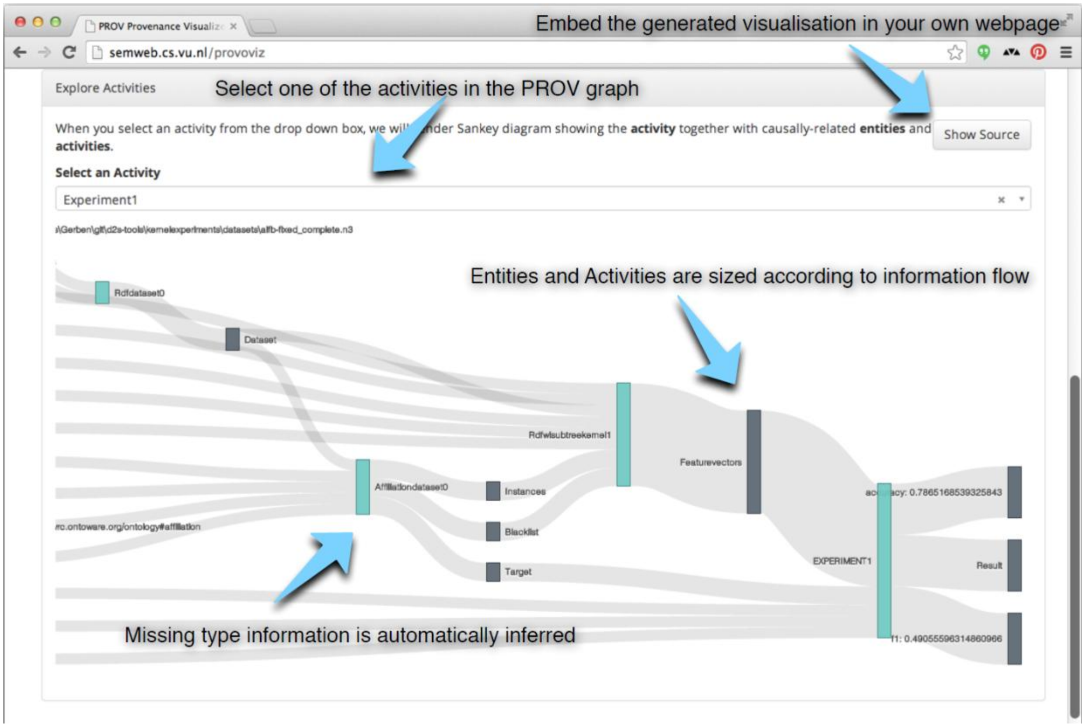


Figure 1: Provenance Map Orbiter : Drill down into summaries [Macko and Seltzer, 2011]

A direct visualization of a subset of the full provenance graph was presented in [Hoekstra and Groth, 2014]. The visualization focuses on the representation of chains of PROV-O activities linked together via the entities used or generated by them via Sankey diagrams. This way, the temporal aspects of data flows within processes can be highlighted, as shown in Figure 2.


Embed the generated visualisation in your own webpage

Select one of the activities in the PROV graph

When you select an activity from the drop down box, we wil nder Sankey diagram showing the **activity** together with causally-related **entities** and **activities**.

Select an Activity

Experiment1

\Gerben\git\d2s-tools\kernelexperiments\datasets\affb-fixed_complete.n3

Entities and Activities are sized according to information flow

Missing type information is automatically inferred

Figure 2: PROV-O-VIZ [Hoekstra and Groth, 2014]

Another direct visualization approach was described in [Kohwalter et al., 2016]. The system allowed the visualization of PROV-O graphs including different means of collapsing and filtering for reducing complexity. One notable feature of the system was its ability to display provenance graphs superimposed on spatial layouts e.g. based on geographical coordinates, which the authors showcased based on a computer game and a bus traffic scenario. The latter is shown in Figure 3, where each dot represents the speed (red-slow, green-fast) of a bus at a given location. Especially the bus traffic scenario, however, blurred the boundary between provenance and actual data.

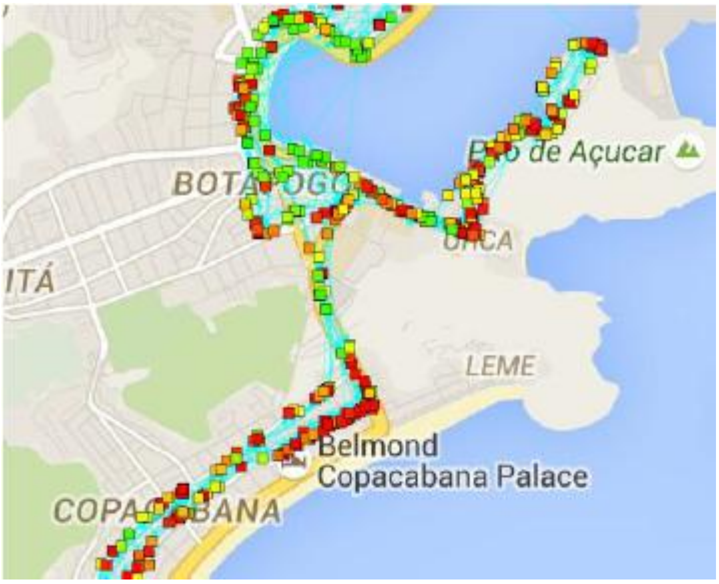


Figure 3: Prov Viewer : Prov on a map [Kohwalter et al., 2016]

VISUALIZATIONS OF PROVENANCE GRAPH ‘DERIVATIVES”

The other group of visualization approaches transforms the provenance graph before it is visualized in order to overcome complexity issues and focus on macroscopic aspects embedded within the recorded provenance information.

One approach to aggregate provenance graph data, “InProv”, was described in [Borkin et al., 2013], focusing on filesystem provenance. The authors proposed a subdivision of the full provenance graph based on the chronological vicinity of the recorded events. Temporally related elements of the provenance graph were grouped together, and for each group, mutual interactions between its member nodes visualized via a dedicated radial layout diagram. Figure 4 shows a screenshot of the proposed system, a separate timeline at the bottom allowed navigation via an overview on the succession and temporal extent of the different temporal clusters, allowing to switch between their respective visualizations. At the right side of the screen, previously selected visualizations were shown in form of a “visual protocol”. Based on the results of a quantitative user study, the authors found that their approach outperformed direct node-link visualizations for larger provenance graphs.

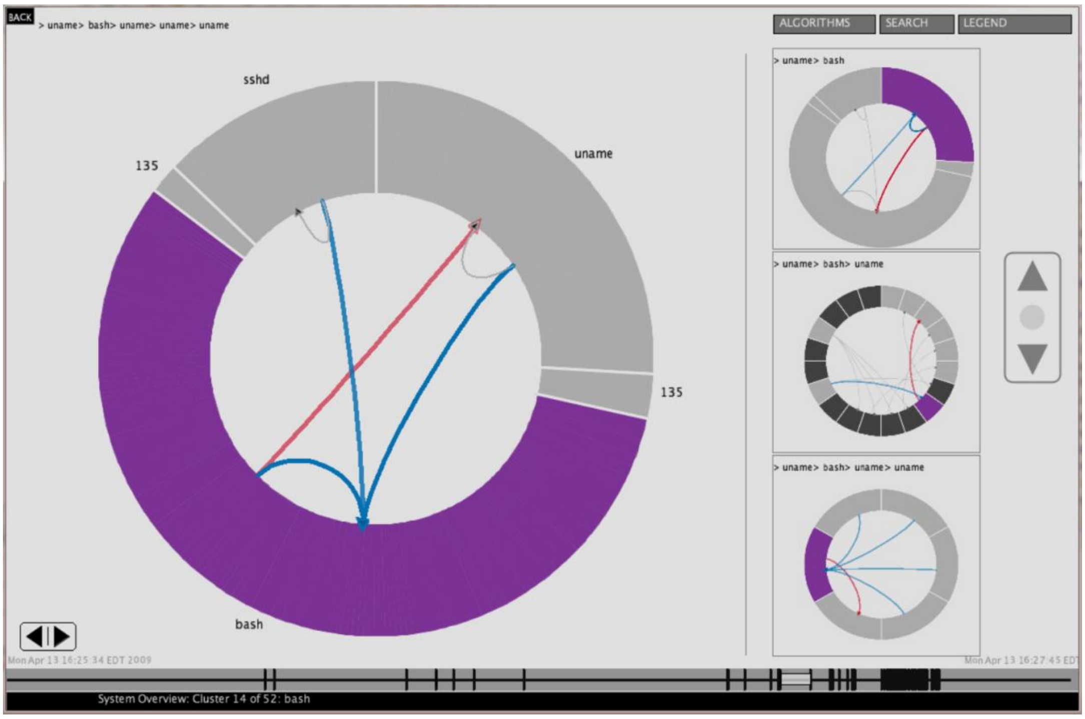

BACK
> uname> bash> uname> uname> uname
ALGORITHMS
SEARCH
LEGEND
sshd
uname
135
135
bash
> uname> bash
> uname> bash> uname
> uname> bash> uname> uname
Mon Apr 13 16:25:34 EDT 2009
Mon Apr 13 16:27:45 EDT
System Overview: Cluster 14 of 52: bash

Figure 4: InProv [Borkin et al., 2013]

As described in [Spinuso et al., 2016], the radial layout approach outlined in [Borkin et al., 2013] was successfully applied to workflow based provenance information in the seismology context. The Bulk Dependency Visualizer (BDV) (see Figure 28) component enabled users to view large scale data dependencies in distributed stream processing environments[20]. It can be used to highlight different provenance perspectives such as data re-use between different users and interactions between different workflows.

2.4 Provenance models

During a session on provenance standardization at the International Provenance and Annotation Workshop (IPAW'06) the first Provenance Challenge on a simple example workflow was set up in order to provide a forum for the community to understand the capabilities of different provenance systems and the expressiveness of their representations (Moreau 2008). After the Third Provenance Challenge, in 2010, the Open Provenance Model (OPM) consolidated itself as de facto standard for representing provenance and was adopted by many workflow systems. The interest of having a standard led to the W3C Provenance Incubator Group, which was followed by the Provenance Working Group. This effort produced the family of PROV specifications[21], which are a set of W3C recommendations on how to model and interchange provenance on the Web.

2.4.1 Open Provenance Model (OPM)

In OPM, provenance is represented by RDF graphs. RDF allows representing the relations (=properties) between the concepts of the model (= classes), based on the triple model. The triple is composed of the subject, the predicate and the object. The subject and the object are

[20] https://github.com/aspinuso/s-provenance

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

individuals of a class, the predicate is the relation that links the two resources. This representation is used to build directed graphs where subjects and objects are nodes and the predicates are directed edges. The directionality of the relation is defined by the domain (D) of the relation, i.e. the class of the origin, and the range (R), i.e. the target of the relation.

OPM is used to describe workflow executions. The nodes in this graph represent three different types of provenance information: resources created as artifacts (immutable pieces of state), steps used as processes (actions or series of actions performed on artifacts) and the entities that control those processes as agents. The edges are directed and have predefined semantic depending on the type of the adjacent nodes: used (a process used some artifact), wasControlledBy (an agent controlled some process), wasGeneratedBy (a process generated an artifact), wasDerivedFrom (an artifact was derived from another artifact) and wasTriggeredBy (a process was triggered by another process). The model is extended by different sub-properties. Roles are used to assign the type of activity that artifacts, processes and agents played in their interaction and accounts are particular views on the provenance of an artifact. Additional information is associated with the relation through the use of the RDF design pattern for named graphs[22], in which a set of RDF statements are identified using an URI. OPM is available as two different ontologies which are built on top of each other: the lightweight OPM Vocabulary (OPMV) and the OPM Ontology (OPMO) with the full functionality of the OPM model. Although this model is superseded by the PROV model it is still used in various workflow environments such as Kepler. OPM does not specify any concept for the modeling of plans, so it can only be used to describe retrospective provenance but not to describe workflow instances or workflow templates.

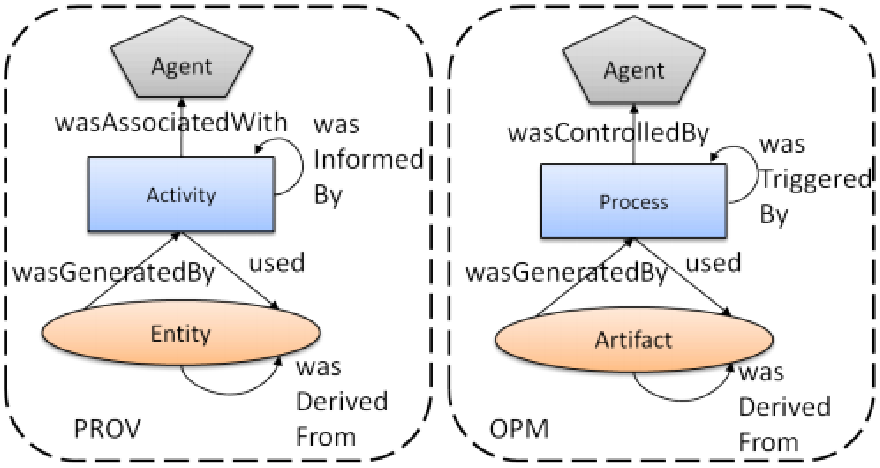


Figure 5: The communalities between PROV (left) and OPM (right) [Garijo 2014].

2.4.2 PROV

The PROV Data Model is very much influenced by OPM, as Garijo [2014] compared and displayed graphically (see Figure 5). This model was released as a standard by the W3C Provenance Working Group in April 2013. PROV is built upon 8 recommendations of the Provenance Incubator Group:

- the core concepts of identifying an object, attributing the object to person or entity, and representing processing steps;

[22] https://en.wikipedia.org/wiki/Named_graph

- accessing provenance-related information expressed in other standards;
- accessing provenance;
- the provenance of provenance;
- reproducibility;
- versioning;
- representing procedures;
- and representing derivation.

PROV consists of 12 documents, which are all needed to understand how to implement provenance. The PROV-Overview[23] helps navigating through the material by classifying each document for a specific reader group (users/developers of applications/advanced). Best is to start with the PROV-PRIMER[24] which provides an introduction to the provenance data model to users by explaining the basic principles via a simple self-contained example. To allow the mapping of the conceptual data model PROV-DM[25] into different representations the Provenance Working Group composed three serializations of it: an ontology (PROV-O)[26], the PROV Extensible Markup Language (PROV-XML)[27] for exchange of provenance data across systems and PROV Notation (PROV-N)[28] which is a specialised notation used to express provenance data in a more human-readable form. All terms defined in PROV have the namespace *http://www.w3.org/ns/prov#* and the prefix *prov*. PROV-CONSTRAINTS[29] is targeted at implementers of provenance validators and PROV-AQ[30] defines how to use Web-based mechanisms to locate and retrieve provenance information. PROV-DC provides a mapping from Dublin Core Terms to the PROV-O. A good overview of existing applications supporting PROV is given in the PROV-IMPLEMENTATIONS report[31], but it has not been updated since 2013. The RDA Working Group of Provenance Patterns[32] plans also to collect tools to assist with the management of provenance which will provide an actual overview of available techniques in near future.

[23] https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
[24] https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/
[25] https://www.w3.org/TR/2013/REC-prov-dm-20130430/
[26] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[27] https://www.w3.org/TR/2013/NOTE-prov-xml-20130430/
[28] https://www.w3.org/TR/2013/NOTE-prov-xml-20130430/
[29] https://www.w3.org/TR/2013/REC-prov-constraints-20130430/
[30] https://www.w3.org/TR/2013/NOTE-prov-aq-20130430/
[31] https://www.w3.org/TR/prov-implementations/
[32] https://www.rd-alliance.org/group/provenance-patterns-wg/case-statement/working-group-provenance-patterns-case-statement

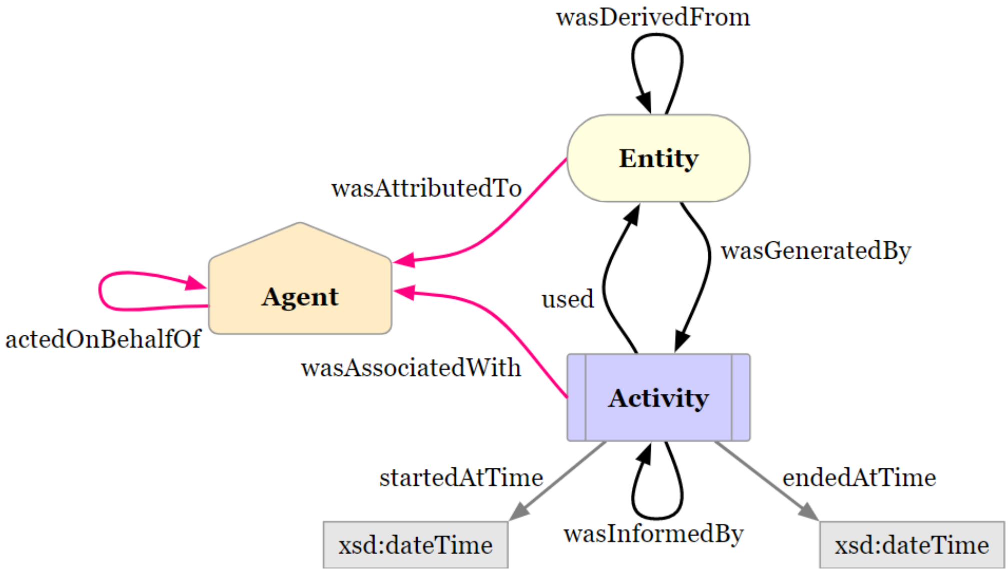


Figure 6: Starting Point Terms

PROV-O expresses the model with OWL 2 and RDF, using the RL (rule-based) profile of OWL 2 to allow reasoning over provenance data. Compared to RDF, the OWL standard goes a step further by distinguishing three main property types: the object property which relates individuals of a class with other individuals (binary relations), the data property which relates individuals with literal data, and the annotation property which relates annotations to OWL elements (class, property, individual, …). Furthermore, OWL allows the axiomatization of the different properties which can be symmetrical, functional and/or transitive.

The *basic elements* of the PROV ontology (http://www.w3.org/ns/prov-o#) 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 classes (compare also to Figure 6):

- *prov:Entity* is the central concept and represents resources (mutable or immutable physical, digital or conceptual things) one wants to provide provenance for. The symbol used to mark Entities is a yellow coloured ellipse.
- *prov:Activity* represents actions performed upon entities such as generating, transforming and modifying one or more entities. It is denoted as a blue rectangle.
- *prov:Agent* is a person or a machine who bears some form of responsibility for an activity. The symbol used for Agents is an orange pentagon.

Following relationships between the classes are modelled:

- as datatype properties:
  - *prov:startedAtTime* and *prov:endedAtTime* [D→*prov:Activity*] - start and end points in time of Activities
- as object properties:
  - *prov:used* [D→*prov:Activity*/R→*prov:Entity]*– an activity used some Entity
  - *prov:wasGeneratedBy* [D→*prov:Entity*/R→*prov:Activity*] - an Activity generated an Entity
  - *prov:wasInformedBy [D→prov:Activity/R→prov:Activity]* **-** an Activity used an entity produced by another Activity, which allows Activity provenance chains
  - *prov:wasDerivedFrom [D→prov:Entity/R→prov:Entity]***-** an Entity was derived from another Entity, without mentioning the Activities involved. This expresses

derivation, the provenance among entities, which is a transformation of an Entity into another.

- *prov:wasAttributedTo* [D→*prov:Entity*/R→*prov:Agent*]– an Entity is ascribed to an Agent
- *prov:wasAssociatedWith* [D→*prov:Activity*/R→*prov:Agent*]– an Activity lies in the responsibility of an Agent
- *prov:actedOnBehalfOf [D→prov:Agent/R→prov:Agent]*- an Agent can be responsible for the Activity of another Agent, who may have been less involved

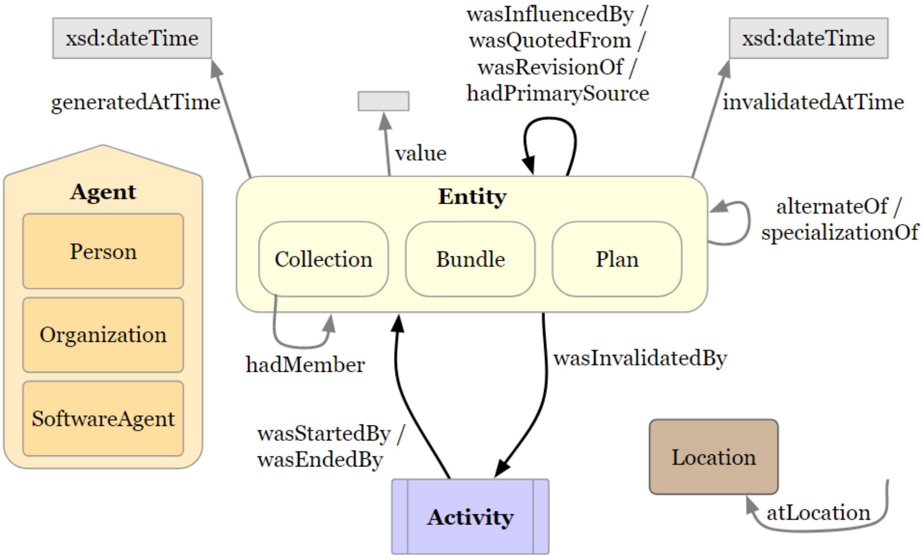


Figure 7: Expanded Terms

Expanded terms of PROV-O encompass among others

- Subclasses of Agent, which may overlap und thus are not disjoint:
  - prov:Person
  - prov:Organization
  - prov:SoftwareAgent
- Subclass of Entity:
  - *prov:Bundle* is a named set of provenance descriptions to allow provenance of provenance
- Datatype properties: to allow time validity descriptions for Activities
  - prov:generatedAtTime
  - prov:invalidatedAtTime
- Subproperties of prov:wasDerivedFrom:
  - *prov:wasQuotedFrom [D→prov:Entity/R→prov:Entity*]- cites a source such as a journal or website
  - *prov:wasRevisionOf* [D→*prov:Entity*/R→*prov:Entity]* - refers to an older version of an Entity

Qualified terms are used to provide additional attributes of the binary relations (object properties). This is an alternative to Named Graphs[33] or RDF reification. The Named Graph approach used in OPM structures the information within an RDF store but does not allow to add this information within the model. The reification of a single RDF triple leads, on the other hand, to the creation of four extra RDF statements because it requires the definition of three parts,

[33] https://en.wikipedia.org/wiki/Named_graph

known as subject, predicate and object in addition to the reification statement itself. This impacts dramatically on query response time. The qualification pattern used in PROV-O is also known as 'qualified relation" which allows creating a class which will represent the relations to which additional information can be associated.

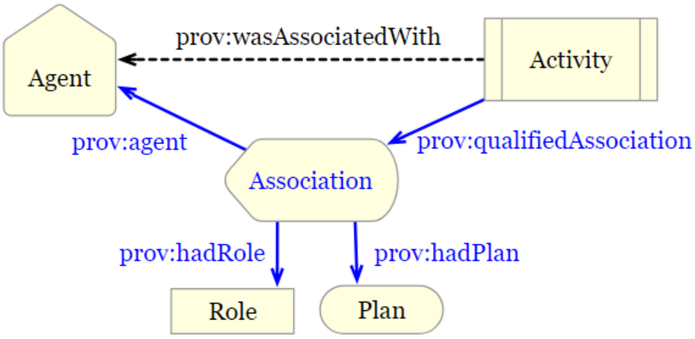


Figure 8: Association as qualified relation class

For example, Figure 8 illustrates that to describe how a *prov:Activity prov:wasAssociatedWith* a particular *prov:Entity*, one creates an instance of *prov:Association* that indicates the influencing entity with the *prov:entity* property. Meanwhile, the influenced *prov:Activity* indicates the *prov:Usage* with the property *prov:qualifiedAssociation*. Now the plan of actions and steps that the Agent used to achieve its goals is provided by adding the object property *prov:hadPlan* to the Association qualified class and an instance of *prov:Plan*. A plan can be a software program, a cooking recipe or anything else that describes how an activity was carried out. Moreover, the *prov:hadRole* property and *prov:Role* class can be used to describe the function that the agent served with respect to the Activity.

Following the pattern described above the following unqualified influence properties (Starting Point relations) can be turned into qualified influence classes to be further described:

- *prov:wasGeneratedBy*→*prov:Generation*
- *prov:wasDerivedBy*→*prov:Derivation*
- *prov:wasAttributedTo*→*prov:Attribution*
- *prov:used*→*prov:Usage*
- *prov:wasInformedBy*→*prov:Communication*
- *prov:actedOnBehalfOf*→*prov:Delegation*
- *prov:wasDerivedBy*→*prov:Derivation*

Similarly the same approach applies with Expanded relations.

In PROV, *time* is defined as datatype properties of Activities (*prov:startedAtTime* and *prov:endedAtTime)* or of Entities *prov:generatedAtTime* and *prov:invalidatedAtTime)*. To be able to add time information to the relationship between Entity and Agent one has to use the qualified classes *prov:Attribution* or *prov:Association*. While PROV is designed to minimize assumptions about time it introduces the concepts of events. These include generation, usage, or invalidation of entities, as well as start or end of activities which are all subclasses of *prov:InstantaneousEvent* that mark transitions in the world. *prov:atTime* is the datatype property to define the time to which an *prov:InstantaneousEvent* occurred.

To describe the provenance of collections, PROV-O provides collection classes and properties as specializations of the Starting Point and Qualified terms. The purpose of this model is to translate domain or application specific provenance representations into a general model. Nevertheless, it is possible to add specific meaning via domain semantics to PROV-O by extending it in a principled way. One can use PROV-O in conjunction with other ontologies by defining direct mappings such us *rdfs:subClassOf* or *rdfs:subPropertyOf* between them. Alternatively it is also possible to extent PROV-O itself at so-called extension points with domain-specific provenance concepts.

2.4.3 Provenir

This is a domain-upper ontology provenance ontology for the translational research domain [Sahoo 2009], developed independently from OPM. It is consistent with other upper ontologies like SUMO (Suggested Upper Merged Ontology), BFO (Basic Formal Ontology) and DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering). Provenir extends primitive philosophical ontology concepts of “continuant” and “occurent” together with ten fundamental relationships stemming from the Relation ontology [Smith et al., 2005].

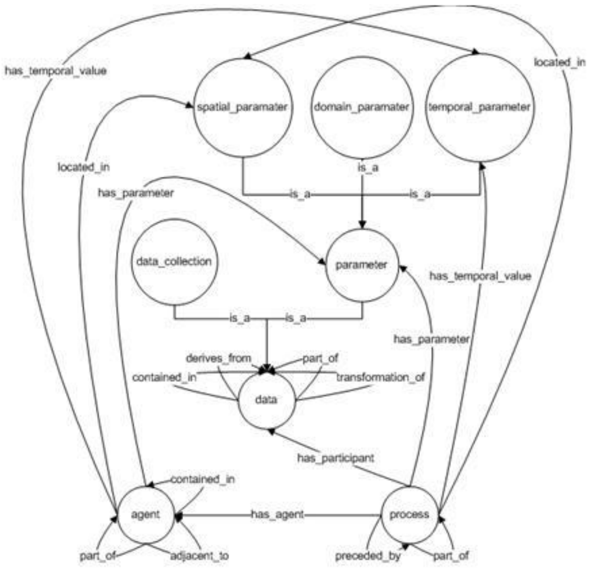


Figure 9: Schema of the Provenir ontology

To represent provenance, three top-level classes are introduced, namely data, process and agent. These are similar to the PROV core classes Entity, Activity and Agent. Data is further specialized in the classes data collection and parameter (spatial, temporal and thematic). Provenir was used as provenance representation in the semantic provenance framework (SPF) to manage provenance information during generation of data from bench experiments and their subsequent use by data mining and knowledge discovery applications. Sahoo et al. [2011] demonstrate that storing provenance and data together based on Provenir allows to store and query provenance information of both pre- and post-publication phase of data. The authors argue that this approach captures the domain specific provenance better than OPM, which lacks also in modelling partonomy, containment and non-causal participatory provenance properties. Patni et al. [2010] use the Provenir ontology to capture and store the provenance of sensor data in their sensor management system.

2.4.4 P-PLAN

In order to be able to represent workflow templates and workflow instances, Garijo and Gil extended PROV [2012]. The plan concept is derived from *prov*:*Plan*, the step concept represents the planned execution activities and the inputs of a step are modelled as a variable with the properties type, restrictions and metadata.

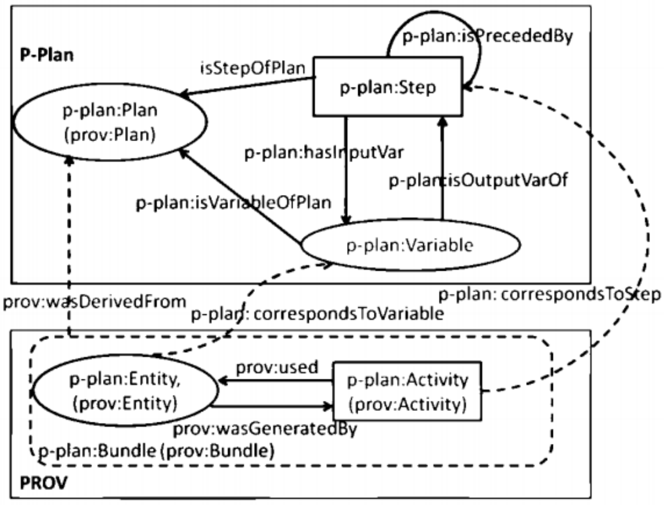


Figure 10: P-Plan [Garijo and Gil, 2012]

2.4.5 OPMW

OPMW is designed to represent scientific workflows at a fine granularity. OPMW extens P-plan, PROV and OPM. It is able to model the links between a workflow template, a workflow instance created from it and a workflow execution that resulted from an instance. Thus, it enables the representation of both r-prov and p-prov. However, Missier et al [2013] criticize, that this is done without introducing any extra vocabulary overloading some OPM terms (e.g. process is used for both provenance representations) and proposes instead the use of ProvONE.

Additionally, it supports the representation of attribution metadata about a workflow reusing terms from the Dublin Core (DC) Metadata Vocabulary[34], namely author, contributor, rights and license.

In OPMW, an *opmw:WorkflowTemplate* is a subclass of *p-plan:Plan* (since it is a particular type of plan), *opmw:WorkflowTemplateProcess* is a subclass of *p-plan:Step* and *opmw:WorkflowTemplateArtifact* extends *p-plan:Variable* respectively [Garijo et al., 2014]. OPMW is used as provenance representation model in the WEST workflow ecosystem.

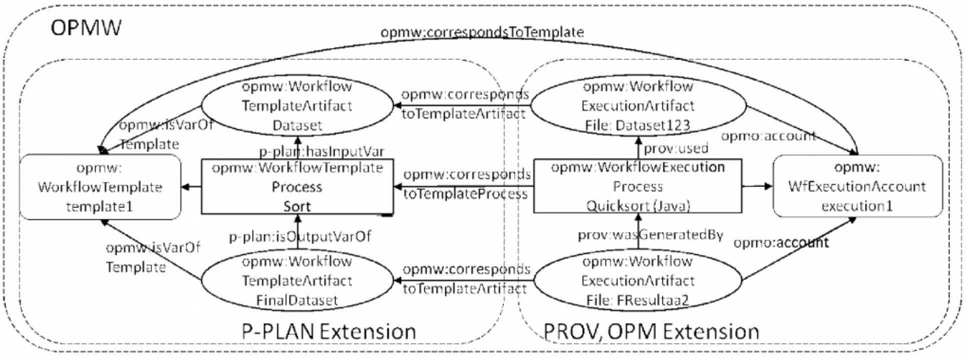


Figure 11: OPMW [Garijo et al., 2014]

[34] http://dublincore.org/documents/dcmi-terms/

2.4.6 ProvONE

Extending and replacing D-PROV model, ProvONE[35], is an ontology that aims to capture all different types of provenance: p-provenance and r-provenance, as well as process provenance and the provenance of data structures [see definitions in 2.2.4].

With PROV it is possible to represent r-prov, using core relations describing that a data item was generated by or was used by an activity which can be a task instance, or invocation [Missier et al., 2013]. While PROV offers the concept $prov{:}Plan$ to refer to an entity that was used by some agent whilst carrying out an activity, it does not capture the graph structure of the dataflow itself. ProvONE extends this baseline provenance pattern.

The conceptual model of ProvONE is presented in Figure 12, below. In this UML diagram, classes involved in the workflow representation are blue, classes involved in the workflow execution/trace representation are yellow/orange, the data structure representation corresponds to the purple classes while the workflow evolution is traced by using the Derivation class of PROV (i.e. the qualified name of the relation “wasDerivedFrom”).

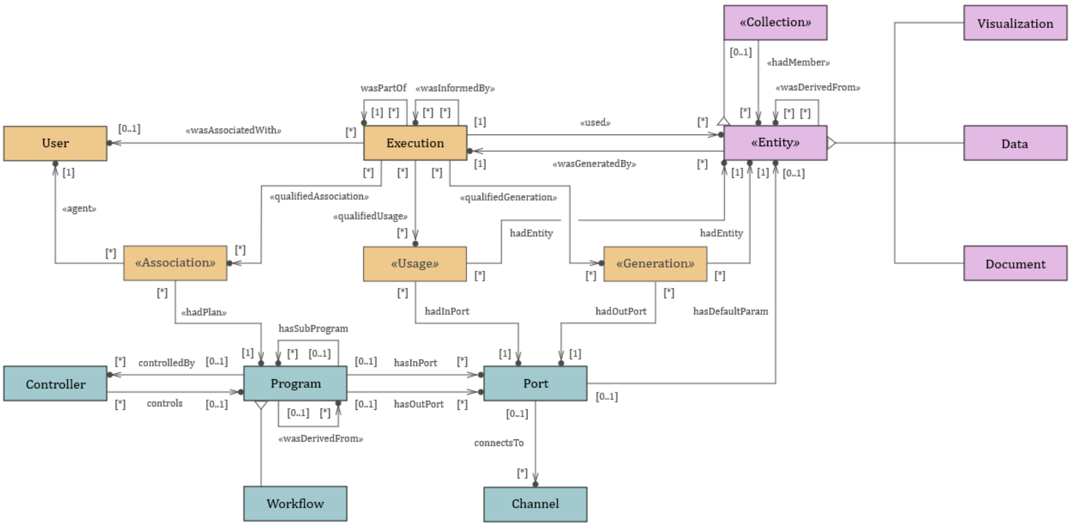


Figure 12: The ProvOne conceptual model UML Diagram

2.4.7 PROV-Wf

The PROV-Wf model is a W3C PROV-DM specialization which is used to represent both p-prov as well as r-prov of scientific workflows that can be provided at runtime. The PROV-Wf model is composed of three main parts: the structure of the experiment (white classes in the UML class diagram), execution of the experiment (dark gray classes) and environment configuration (light gray classes) [Costa et al., 2013].

[35] http://tinyurl.com/ProvOne

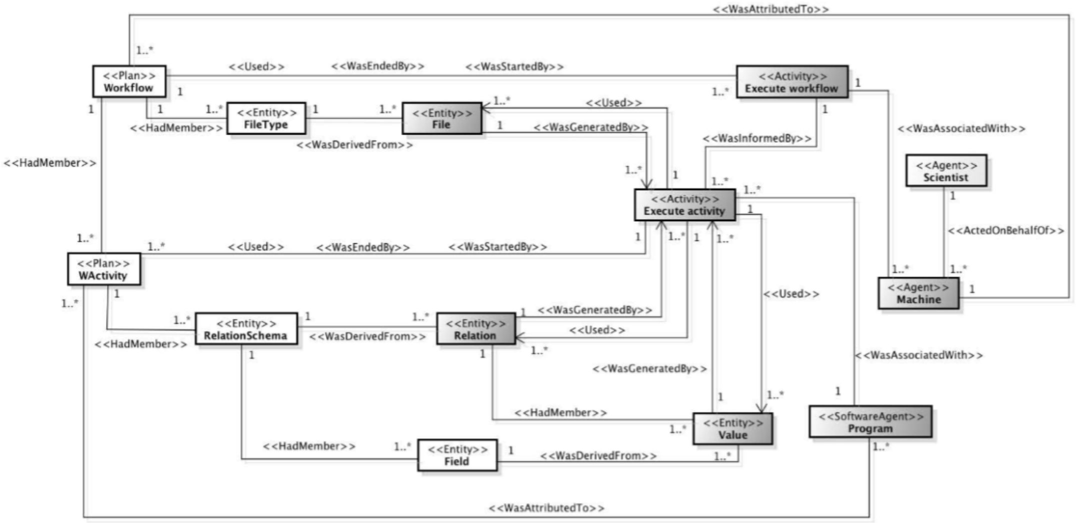


Figure 13: PROV-Wf data model [Costa et al., 2013]

The agent *Scientist* uses computational resources to execute the experiment (composed as a workflow). Furthermore, the Scientist is associated to a machine (i.e. agent *Machine*). The *Machine* establishes an association with a scientific workflow (plan *Workflow*), which is composed by a set of activities (i.e. plan *WActivity*). Each activity is responsible for executing a program. The invocation of a program within a workflow (i.e. *Execute Activity*) uses a set of parameters (i.e. *Field*) that can be seen as a set of values to be consumed. To express all data that are consumed and produced by execution instances, the entity Relation is associated with a schema which can be defined with multiple fields. Each field (i.e. entity *Field*) describes the meaning of each parameter associated to a program that is associated to a *WActivity*. The entity *Value* expresses the set of values of a field, each set associated to an execution instance. Furthermore, the entity *File* represents all files consumed and produced by a workflow execution and entity *FileType* represents the expected type of file to workflow. Associations of Entities are expressed by *Used* (consumption) and *WasGeneratedBy* (production). Activities are modelled as action that happens in a period of time (during workflow execution) using some entities and a plan.

An extension of PROV-Wf adds some classes to this model with the purpose of supporting domain data provenance. According to De Oliviera et. al. [2015], this increases the power of provenance data analysis which depends on the expressiveness of domain-specific data along the provenance traces. FileType is replaced by ValueType representing a specific data type or structure that can be a File or domain-specific data (class Domain Data). In order to model the collection of domain-specific data from produced raw data files, the class Execute Extractor is introduced. This program can be triggered by an activity execution. This model allows to query domain data values related to other provenance data.

2.4.8 S-PROV

S-PROV[36] utilises and extends PROV and ProvONE models. It allows users to analyse the relationships characterising the computational methods at different levels of granularity and detail. The model addresses aspects of mapping between logical representation and concrete

[36] https://github.com/KNMI/s-provenance

implementation of a workflow until its enactment onto a target computational resource. The model captures aspects associated with the distribution of the computation, volatile and materialised data-flow and the management of the internal state of each concrete process. It tracks runtime changes, especially in support of flexible metadata management and discovery for the data products generated by the execution of a data-intensive workflow. It serves as underlying model for S-ProvFlow, a provenance framework for storage and access of data-intensive streaming lineage (see for me detail chapter 4.2.3). This framework is integrated within the VERCE Earthquakes Simulation portal[37] and the climate services portal[38].

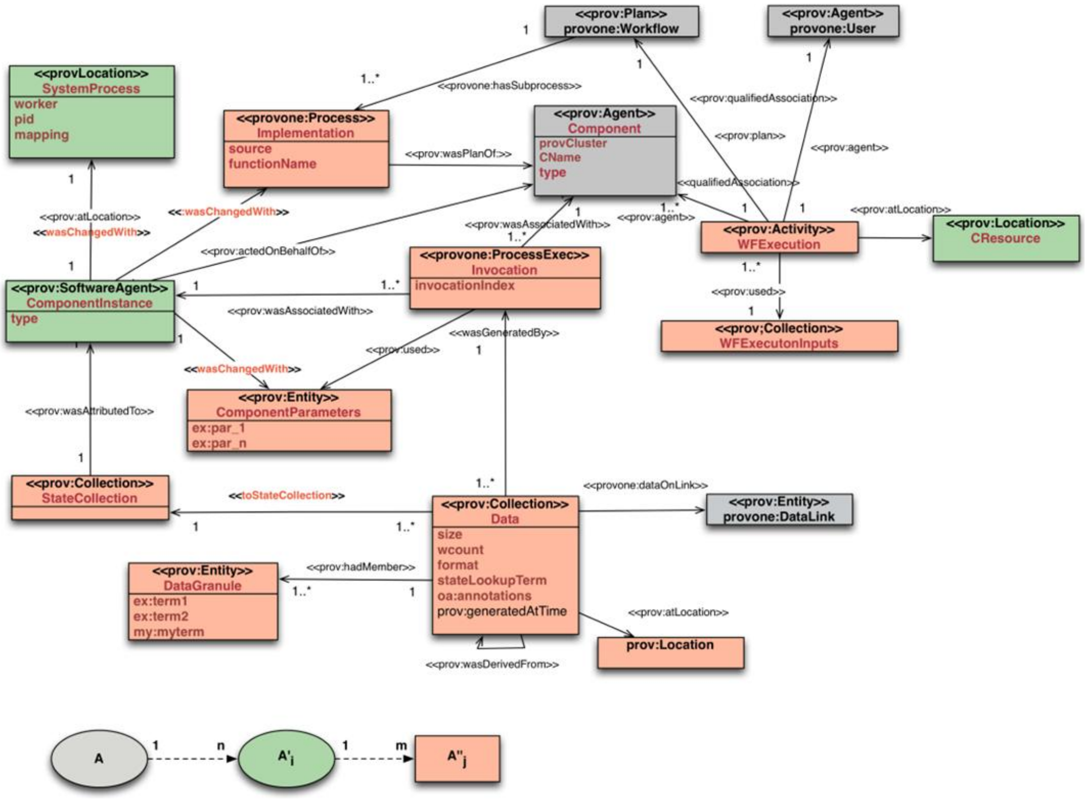


Figure 14: S-PROV Model - abstract definition (grey), concrete deployment (green), runtime retrospective (red) [Spinuso, 2018a]

2.4.9 W4Ever models

The Wf4Ever Models provide a vocabulary for the description of Research Objects[39] that are workflow centric.

[37] http://portal.verce.eu

[38] http://climate4impact.eu

[39] http://wf4ever.github.io/ro/2016-01-28/

RESEARCH OBJECT MODEL

Research Objects (RO) are defined as self-contained units of knowledge, facilitating the publication, sharing and reuse of data. The Core Principle of ROs are *identity*, using global unique identifiers, *aggregation* to associate things that are related or part of the broader investigation and *annotation* providing additional metadata for better discoverability. A research object aggregates a number of resources that are used and/or produced in a given scientific investigation. This aggregation provides access to those collected resources – or at least access to the identification of those resources. The model allows for annotation of the aggregated resources, offering a container within which these annotations can be asserted, allowing for the capture of context and provenance concerning those relations and annotations (see Figure 15).

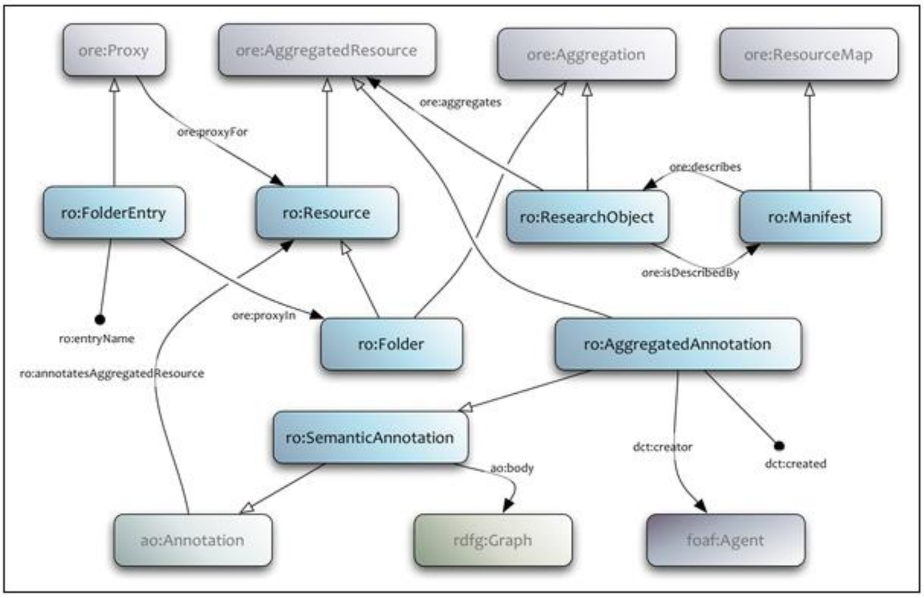


Figure 15: Research Object Model

Although the basic infrastructure (aggregation + annotation + domain vocabularies) that the RO model supports is applicable to many situations the focus of Wf4Ever Models is on aggregations of resources relating to scientific workflows.

The RO model is an extension of the concept of workflow bundle proposed in SCULF2, which is the mechanism used to specify Taverna workflows. Taverna was created by the myGrid team[40] and is now an Apache Incubator project[41]. It is used in the life sciences communities in connection with the workflow repository platform myExperiment[42] which allows scientists to share and reuse workflows. SCULF2 is based on Linked Data technology and defined using the SCULF2 OWL Ontology and annotation with URIs which allows third parties to extend the workflow annotation with additional information. Although SCULF2 is designed to describe workflows, the modelling and the description of provenance needs to be tackled separately.

The intention is that a Wf4Ever Research Object aggregate information about a workflow including details of its execution trace, the data items consumed or produced by the workflow, plus provenance information about the lineage of the workflow, data items or the aggregation

[40] http://www.mygrid.org.uk/
[41] https://taverna.incubator.apache.org/
[42] https://www.myexperiment.org/home

itself. This supports reproducibility of workflows, reuse of the components and, perhaps most importantly facilitates subsequent understanding of the investigation that the workflow is intended to support.

Missier et al. [2013] emphasizes that D-PROV (thus also ProvONE because it is an extension of D-PROV) is fundamentally different from the Wf4Ever models because the first acts as a global model whereas the latter are confined to data-driven workflows. While D-PROV supports both channel- and port-based workflows, wfdesc and wfprov are limited to port-based workflows. Moreover D-PROV is suitable for executable workflows with implemented steps by software components while the Wf4ever models can additional be used to document abstract workflows.

Wf4ever Research Object model comprises three main ontologies to support workflows[43].

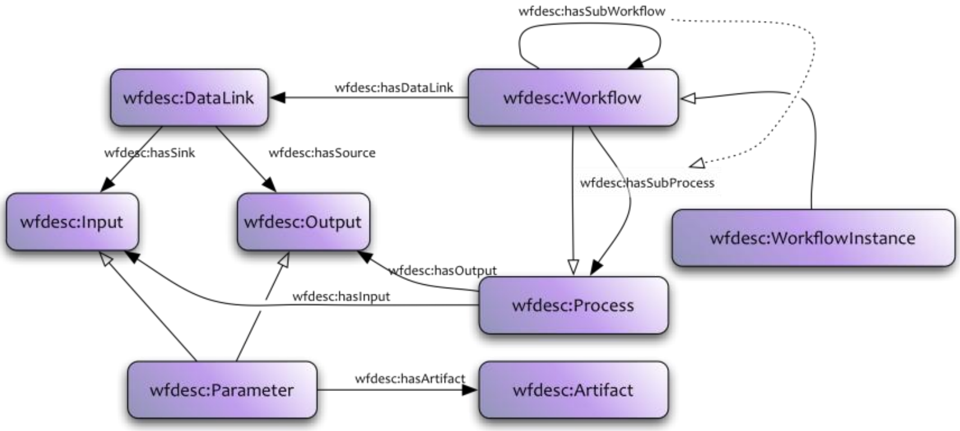


Figure 16: wfdesc ontology

[43] http://wf4ever.github.io/ro/#ontologies

The wfdesc ontology is a vocabulary for the description of workflows. It is an upper ontology to align more specific workflow definitions and to express abstract workflow from SCULF2 and others, capable to describe p-prov.

The wfprov ontology is designed to describe workflow execution provenance trace in alignment with wfdesc, thus specifying the r-prov. It provides an abstraction that can be mapped to different particular workflow systems.

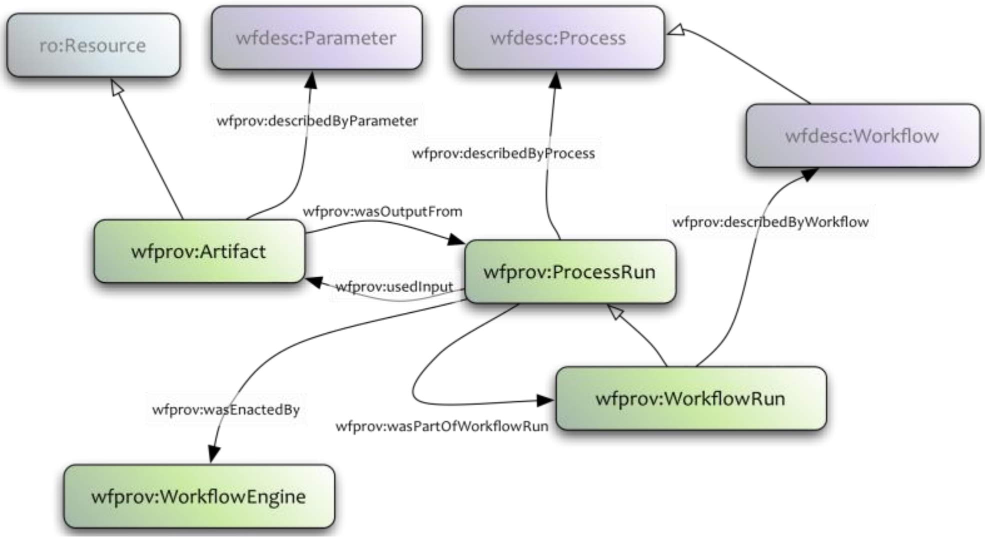


Figure 17: wfprov ontology

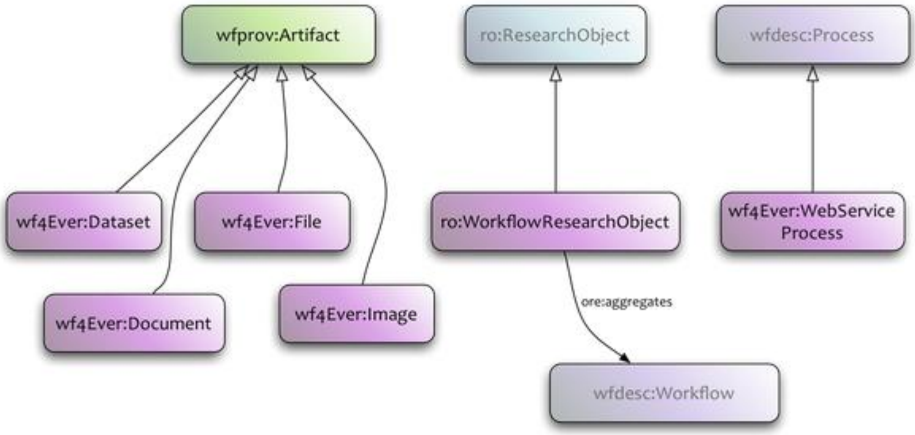


Figure 18: wfever ontology

The Wf4Ever ontology defines extensions specific to Wf4Ever and aggregate the other WF4Ever ontologies.

This representation of provenance is generic and gives the possibility to produce W3C PROV description by using a specific plugin. Nevertheless, the concepts used to build this model are not aligned directly with the PROV model, necessitating therefore a mapping effort.

2.4.10 CERIF

CERIF (the Common European Research Information Format) is a formal conceptual model describing the research domain. It supports the management of Research Information, which is information about research entities such as people, projects, organisations, publication, products, etc. and the relationships between them.

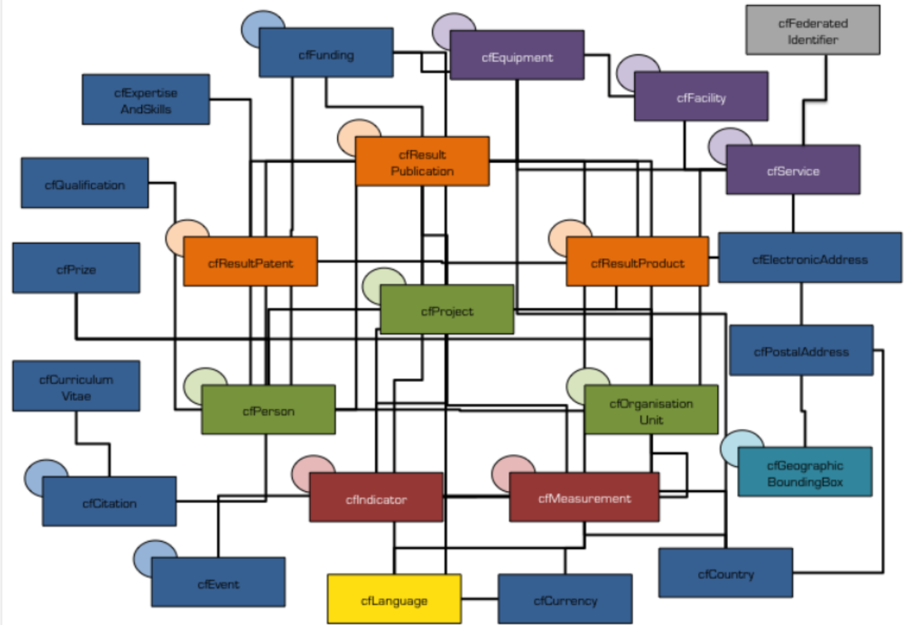


Figure 19: CERIF entities and their relationships (in green: actors, in orange: results, in red: outcomes, in purple: infrastructure...)

CERIF is described as an extended entity-relationship model with temporal additions. It supports the encoding of ontological relationships between concepts generally represented as tables in the model. CERIF's syntactic layer includes the concepts of base entities representing things (entities or objects in the real world) and relationship or linking entities (which express relationships between base entities). Each entity is provided with a system-internal identifier and some basic entity-specific attributes. Linking entities consist of links to the two concerned base entities and the role and temporal duration of the relationship. The role is a semantic identifier, referencing the term, description and relationships to other terms (e.g. multilingual, crosswalking, thesaural and ontological relationships) in the CERIF semantic layer.

The CERIF team developed a mapping CERIF-RDF for Linked open data already starting 2008 and specifically to link European CERIF CRIS (Current Research Information Systems) to the US VIVO system for universities. A specific LOD group was formed in 2011. The generic work on CERIF RDF was used in the ENGAGE project to generate the ENGAGE metadata format for the portal for general user interface requirements [Zuiderwijk et al., 2013]. The ENGAGE infrastructure uses a three-layer structure for metadata and CERIF was used for the implementation of the middle, contextual, layer because it offers temporally defined role-based relationships between instances of entities.

Because of the time-stamped linking entities, instances in a relationship provide provenance records (as well as constraint records and linking or relationship records). PROV-O attaches start and end times to activities and time stamps to when certain things happen, but not as universally to all kinds of roles and events as CERIF enables. Introducing the temporal modelling with the qualification pattern in PROV-O requires many triples to represent one CERIF n-tuple. This has performance and storage implications. Compton et al [2014] demonstrated that by mapping CERIF to PROV-O the provenance of research results could be significantly enriched compared to using solely PROV-O. Nevertheless, according to [Bailo et al., 2016] there is still necessity to further develop in CERIF some provenance aspects such as the integration of causal-effect relationships among the entities and activities involved and re-used across processing tasks.

2.4.11 Blockchain

As a by-product property of their design, blockchains provide provenance (rather like transaction logs in database systems but within the blockchain). Blockchain is proposed as a solution to provide trust especially when data lineage about tasks and processes in distributed system are collected and stored by many agents at the same time [Stoffers, 2017]. Proof of integrity is only given when data remained unchanged since they were stored or processed. By combining peer-peer networking with Merkle trees, asymmetric cryptography, time-stamping and proof-of-work, Nakomoto [2008] developed a tamper-resistant distributed database for maintaining ownership of digital money, including the complete lineage of all occurred transactions. However, to utilise the provenance characteristics of blockchains the application has to be implemented using blockchain technology. It is not an 'add-on' like, for example, PROV nor part of an integrated catalogue solution.

2.4.12 CKAN

CKAN is a tool for making open data websites and is widely used by national and local governments and research institutions. It helps to manage and publish data in units called datasets. It supports users by providing faceted search features to browse and find datasets they are interested in. Datasets contain a rich set of metadata about the data (like title, identifier, description, ...) and a number of resources holding the data themselves in any format the data are available. The CKAN metadata used the Open Knowledge Foundation's Version Domain Model (VDM)[44] to keep a complete history of all edits and version of the dataset metadata. However, this feature has been deprecated because it was not supported. It is being replaced by an improved Activity Stream for a dataset which links to view each version as it was, providing some sort of provenance information[45].

2.5 Provenance related initiatives

2.5.1 The DataONE perspective

Data Observation Network for Earth (DataONE)[46] is the foundation of new innovative environmental science through a distributed framework and sustainable cyberinfrastructure that meets the needs of science and society for open, persistent, robust, and secure access to well-described and easily discovered Earth observational data.

The DataONE ProvWG (Provenance in scientific workflows group) lead by B. Ludaescher and P. Missier, is developing an open and extensible provenance management architecture for scientific data processing systems.

[44] https://pythonhosted.org/vdm/

[45] https://github.com/ckan/ckan/pull/3972

[46] https://www.dataone.org/

During phase I of DataONE, the group did extensive work on interoperable provenance across workflow systems, notably Kepler and Taverna, by developing:

- provenance (data) model (D-OPM), and the follow-up ProvONE model
- new first-class objects in DataONE (Figures, Software)
- an Extended Data Package with Provenance

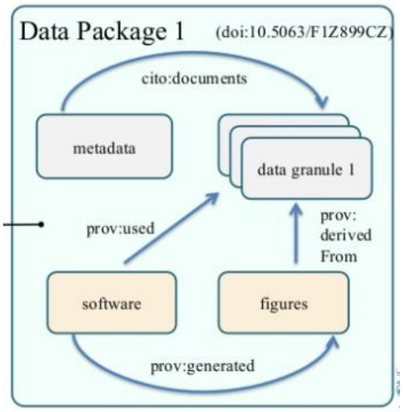
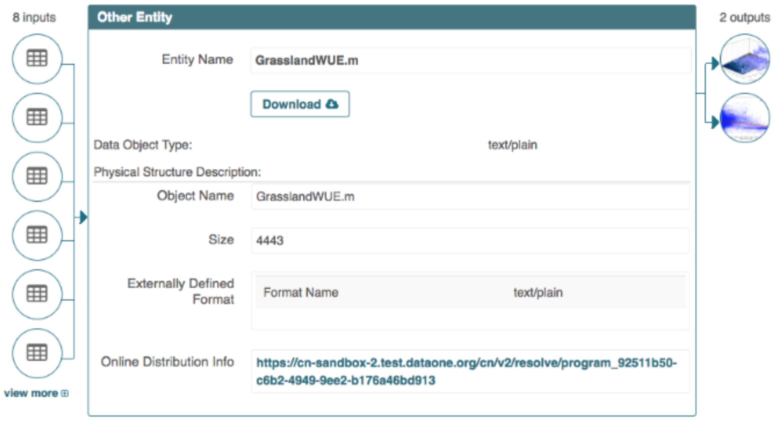


Figure 20: Extended Data Package

During phase II, the WG collected provenance use cases and related requirements[47]. These will be ingested in the RDA Provenance Patterns Database. The group also developed a number of tools, such as provenance indexing, Matlab tool for generating provenance, R tool for generating provenance, YesWorkflow tool for modeling provenance from scripts.

A very promising implementation of provenance display is the Provenance explorer. This is a graphical web interface to explore provenance of generated products in the DataOne portal based on the Extended Data Package which enables the storage of provenance information at the level of data granules composing the dataset.



Other Entity

Entity Name GrasslandWUE.m

Data Object Type: text/plain

Physical Structure Description:

Object Name GrasslandWUE.m

Size 4443

Externally Defined Format — Format Name text/plain

Online Distribution Info https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_92511b50-c6b2-4949-9ee2-b176a46bd913

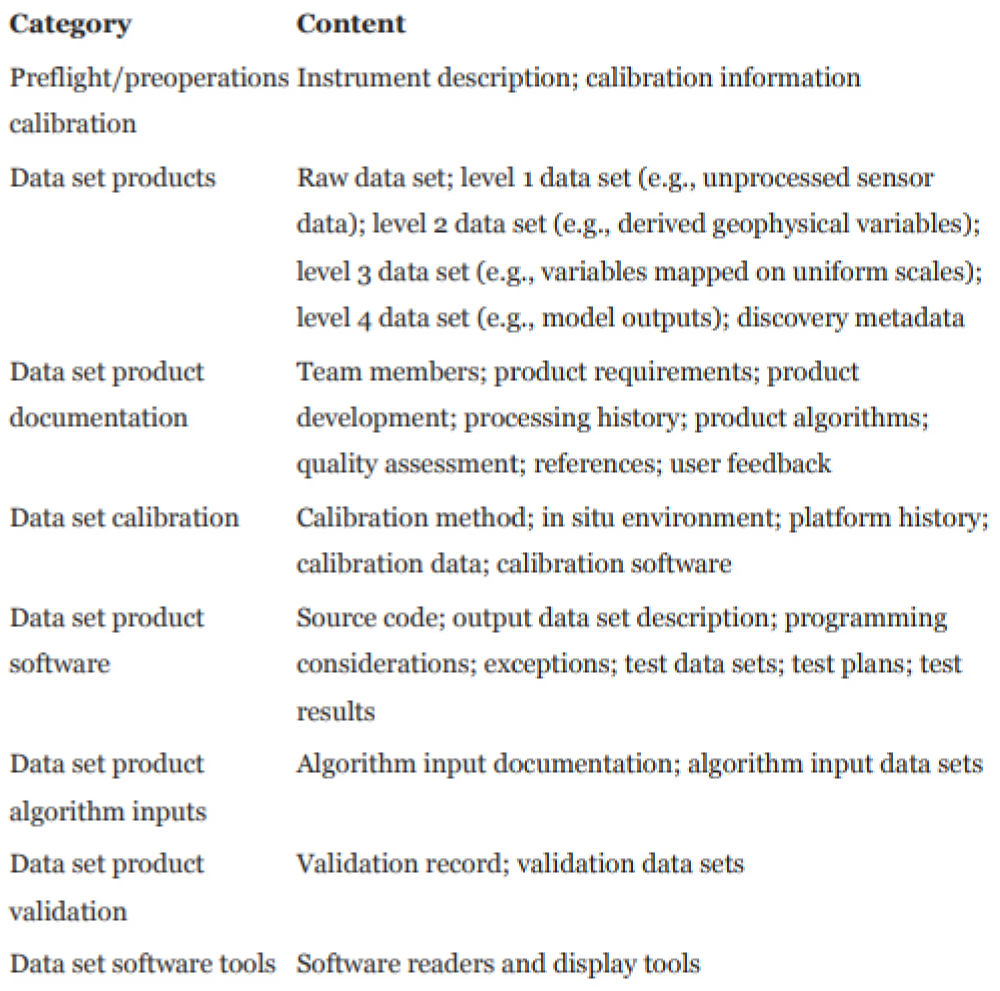
Figure 21: Provenance explorer

[47] https://docs.google.com/spreadsheets/d/1r0qeE7RgfYtC4Y-rGkxfFKezeyc8rRMcZ3wdn8zYZOE/edit#gid=0

2.5.2 The ESIP perspective

Earth Science Information Partners (ESIP) is a non-profit, volunteer and community-driven organization that advances the use of Earth science data.[48] ESIPs activities are organized around three types of collaboration areas: Committees, Working Groups and Clusters. The Data Stewardship Committee develops and fosters practices and standards in the field of Earth science informatics aiming to facilitate their long-term management, preservation and curation.[49] One activity is about provenance which aims to produce a provenance and context content standard[50].

Downs et al. [2015] demand that journals must require that data are not only available but also reusable. They recommend new approaches to identify, capture, and track all details necessary to demonstrate data validity and to better ensure scientific reproducibility, through the Provenance and Context Content Standard (PCCS) matrix. The matrix which details the content required to describe provenance and context and defines eight categories of data, metadata and documentation that need to be preserved and accessible.



Category	**Content**
Preflight/preoperations calibration	Instrument description; calibration information
Data set products	Raw data set; level 1 data set (e.g., unprocessed sensor data); level 2 data set (e.g., derived geophysical variables); level 3 data set (e.g., variables mapped on uniform scales); level 4 data set (e.g., model outputs); discovery metadata
Data set product documentation	Team members; product requirements; product development; processing history; product algorithms; quality assessment; references; user feedback
Data set calibration	Calibration method; in situ environment; platform history; calibration data; calibration software
Data set product software	Source code; output data set description; programming considerations; exceptions; test data sets; test plans; test results
Data set product algorithm inputs	Algorithm input documentation; algorithm input data sets
Data set product validation	Validation record; validation data sets
Data set software tools	Software readers and display tools

Figure 22: The Provenance and Context Content Standard (PCCS) Matrix [Downs et al., 2015]

[48] http://www.esipfed.org/about

[49] http://wiki.esipfed.org/index.php/Preservation_and_Stewardship

[50] http://wiki.esipfed.org/index.php/Provenance_and_Context_Content_Standard

PCCS has been adopted by NASA within the Earth Science Data Preservation Content Specification[51].

Within ESIP an Information Quality Cluster[52] was established in 2015 with H.K. Ramapriyan as chair. Its vision is to become an authoritative and responsive resource of information and guidance to data providers on how to implement data quality standards and best practices including provenance aspects. It is still active and open to all, organizing monthly teleconferences (Second Tuesdays at 11 ET) and organizing sessions in international conferences (AGU, RDA).

2.5.3 The RDA perspective

The Research Data Alliance (RDA)[53], launched in 2013, is an international organization focused on the development of infrastructure and community activities aimed to reduce barriers to data sharing and exchange, and promote the acceleration of data driven innovation worldwide. With more than 6,700 members representing 136 countries, RDA includes researchers, scientists and data science professionals working in multiple disciplines, domains and thematic fields and from different types of organisations across the globe.

RDA Interest Groups (IG)[54] are comprised of experts from the community that are committed to directly or indirectly enabling data sharing, exchange, or interoperability. These groups must have international participation and a demonstrated community and should not be for promoting specific projects or technologies. They are long-term initiatives within the RDA and remain in operation as long as they remain active. They serve as a platform for communication and coordination among individuals, outside and within RDA, with shared interests. They produce important deliverables such as surveys, recommendations, reports, and Working Group case statements.

RDA Working Groups (WGs)[55] should tangibly accelerate progress in concrete ways for specific communities with the overarching goal of increasing data-driven innovation. They should strive for harvestable efforts within a lifetime of 12-18 months that have substantive applicability to particular segments of the data community.

There are several groups within RDA dealing with provenance issues. The reason for this is that provenance is foundational to many other RDA groups' activities.

RDA groups having aspects included but not focusing specifically on provenance:

- Publishing Data Workflows WG (already completed)
- Dynamic Data Citation WG (already completed)
- PID Kernel Information Types WG (already completed)
- Reproducibility IG (active)
- PID IG (active)
- Archives and records professionals for research data IG (active)

[51] (https://earthdata.nasa.gov/standards/preservation-content-spec)

[52] http://wiki.esipfed.org/index.php/Information_Quality

[53] https://www.rd-alliance.org

[54] https://www.rd-alliance.org/groups/creating-and-managing-rda-groups/creating-or-joining-rda-interest-group.html

[55] https://www.rd-alliance.org/groups/creating-and-managing-rda-groups/creating-or-joining-rda-working-group.html

- Data discovery paradigms IG (active)
- Preservation e-Infrastructure IG (active)
- From Observational Data to Information IG (active)
- Metadata IG (active)
- Data in Context IG (active)

There is an active IG and an active WG specifically addressing provenance lead by Nicholas Car, Dave Dubin and Paolo Missier, in which several authors of this deliverable are also involved in:

- Research Data Provenance IG
- Provenance Pattern WG (PPWG)

The IG focuses on the comparison and evaluation of models for data provenance. It is concerned with questions of data origins, maintenance of identity through the data lifecycle, and how we account for data modification. Objectives of this group include: recommending general and expressive frameworks for documenting research data transactions proposing syntheses of complementary provenance views, and relating data provenance to problems of scientific equivalence and the assessment of data quality. The activities of the IG are in the moment waiting for input from the WG.

The WG was started in September 2017 and meets regularly twice a month (one telco at Austrialian friendly time - Wednesday 1am UTC, and one telco at European friendly time – Tuesday 2pm UTC). It seeks to help science data communities to adopt existing provenance management practice. It aims to find, detail and recommend best practices for provenance representation and management. It looks for existing practice rather than generate new practice.

The Provenance Patterns WG has two activity areas:

1. Common provenance Use Cases (UC) (see chapter 3.2.2)
2. Provenance design patterns (PP)(see chapter 4.1)

The WG will engage with the other RDA groups listed above and source its primary requirements and exemplars from these.

3 ENVRIPLUS PROVENANCE REQUIREMENTS

3.1 Methodology

As stated in the introduction it is impossible to design a one-size-fits-all system for all domain and application areas. We thus focus on collecting explicit requirements from RIs and provide them advice how to approach individual implementations of provenance management systems that fit their architectures. Existing tools can generally be used but have to be adapted to the technology choices made within their architectures. More importantly it is helpful to find the right strategy how provenance should be introduced in the existing infrastructures and RIs should thus be provided with guidance on key modelling and encoding decisions.

Fortunately there is already a widely used and acknowledged standard for provenance (W3C – PROV documents) we can rely on. The PROV ontology provides a generic model for implementing provenance applications that can represent, exchange and integrate provenance information generated in different systems and under different contexts. Being domain-agnostic, it must usually be refined for specific application areas and RI objectives. The ontology provides means for extending the general model but it is not specified in the documentation how this should be done. Many groups and initiatives (ESIP, DataONE, RDA, see also chapter 2.5), however, have already dealt with these specifics. The ENVRIplus provenance group takes such existing considerations into account. In particular, we follow the ongoing work of the RDA Provenance Patterns WG (PPWG) which aims at collecting provenance Use Cases and related Provenance Patterns to propose best practices.

For the first phase (October 2017 – April 2018), the approach of the ENVRIplus provenance group was structured as follows:

- Collect requirements and use case of the ENVRI RIs
- Compare their descriptions and identify common ones, if possible generalize them
- Compare descriptions with those provided by other initiatives, in particular:
    - the requirements with those of the Provenance Incubator Group
    - the use cases with those of the RDA Provenance Patterns Database (PPD)
- Add ENVRI use cases to the PPD if not yet included

In the second phase (May - October 2018) we envisage to:

- model selected ENVRIplus use cases as
    - activity diagrams in UML
    - transcribe them as OIL-E instantiations
- associate RDA Provenance Patterns with the ENVRIplus use cases if available (the RDA PPWG has only produced a few patterns so far but is going to develop further ones in the coming year which should provide advice for the collected use cases in the database)
- for selected RDA Provenance Patterns
    - model them as activity diagrams in UML
    - transcribe them as OIL-E instantiations
    - optionally: transcribe these in terms of CERIF to have a mean to compare the different approaches
    - incorporate them in the ENVRI Knowledge Base to make them available as best practice, accessible via a demonstration platform for interested parties [Martin, 2018]

Details regarding the second phase still have to be confirmed by the ENVRIplus provenance group at the ENVRIweek in May 2018. Moreover we intend to demonstrate how provenance can be implemented on specific use cases (such as TC17 see chapter 5.2). There may be additional requests from the Ris, e.g. to test provenance tools and services.

3.2 Requirements and use cases background

3.2.1 Definitions

USE CASES

Use cases are a description of a set of interactions between a system and one or more actors. They describe details of a function in the system. They can be seen as user-perspective specifications. For each use case one might find more than one requirement. This dependency should be stored together with the requirement.

REQUIREMENTS

By requirement we mean a formal description of what a user expects from the system. We distinguish:

1. Functional - requirements that are translated to use cases or user stories and then implemented in business logic of the application.
2. Non-functional - requirements that (in most cases) should be ensured by the architecture of the system. These are aspects of use of the requirement such as operational or usability and performance, which can be added if considered relevant.

Requirements have a functional perspective, they approach the problem from the angle of a solution. They can be understood as developer specifications. Requirements should be mapped to use cases if possible.

3.2.2 RDA Provenance Use Cases

The first activity of the RDA Provenance Patterns WG (PPWG) is to collect and document provenance use cases (UC). A provenance use case recording system was implemented[56] for this purpose.

The use cases are provided by the members of the WG or sought from other IG/WGs.

The basic elements of the UC which should be reported are: contributor, actors, broader UC, goal, summary, steps, alternative steps (see test UC[57]).

The PPWG aims at generalizing the list of collected Use Cases by de-duplication and categorisation in order to reveal common features and structure, since many provenance use cases are "just" instances of general cases, separated by different domain terminologies.

The establishment of a published set of UCs will allow people to compare their UCs with known UCs for which recommended implementations and other patterns may already be known. It will also allow people to consider provenance UCs posed by others that may be of future interest to them.

The UCs collected so far (assessed: April 2018) are not yet generalized and in many cases overlap. This will be improved in the upcoming months (2018) by the RDA PPWG as explained above.

3.2.3 Provenance Incubator Group requirements

The mission of the Provenance Incubator Group, formed in 2009, was to provide a state-of-the-art understanding and to develop a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization. This group did the basic research that lead to the PROV specifications. They produced a number of documents, including:

- an overview of key dimensions for provenance
- more than thirty use cases spanning many areas and contexts, illustrating these key dimensions
- a broad set of user requirements and technical requirements[58] derived from those use cases.

The requirements are organized around the identified key dimensions. The list is very comprehensive and a good source for comparison.

[56] http://patterns.promsns.org

[57] http://patterns.promsns.org/usecase/62

[58] https://www.w3.org/2005/Incubator/prov/wiki/Requirements

3.2.4 Process for requirements and use case collection

At the beginning of the development of any RI system design it is important to understand the existing situation and conditions of the RI architecture related to technology of interest. A good way to study this is the collection of requirements and use cases via RI representatives who have good insights in the RI architecture.

3.2.4.1 First requirements gathering (2015/2016)

In conjunction with the task 5.1 “Review with existing Ris” the provenance team participated in the requirements study to analyse the status of involved RIs regarding provenance management and implementation. This process involved three roles:

- The role of a topic leader: They had to be receptive to input from ICT-RI go-betweens and had to partition and delimit their topic to minimise duplication of work by those contributing to their topic.
- The role of an RI representative (RIREP) was to collect and present to requirement gatherers information about their RI's requirements, including its existing inventory of facilities, its plans as they affect technical choices, their alliances with e-Infrastructure providers and the work of various roles within their RI who need better data facilities. They introduced other members from their RI into the requirements gathering process to work directly on specific issues or topics.
- The role of an ICT-RI go-between (GB) was to avoid duplication of effort by an RIREP in an RI they are responsible for. The GBs were guided by a common set of information requirements.

The provenance related section of the requirements questionnaire intended to collect whether provenance was already considered in each RI's data lifecycle and if any related implementations were already in use. For those Ris where the latter was not the case, the next set of questions was focused on their potential interest in provenance tracking: which type of information should be tracked, which standards to rely on and finally which sort of support was expected from ENVRIplus. Most RIs already considered provenance data as essential and were interested in using a provenance recording system.

Among the nine RIs who gave feedback about provenance, only two already have a data provenance recording system embedded in their data processing workflows. EPOS uses the dispel4py workflow engine in VERCE. IS-ENES2 instead did not directly specify their applied software solution. Some RIs, such as SeaDataNet and Euro-ARGO, interpret provenance as lineage metadata gathered with tools (Geonetwork) based on metadata standards such as ISO19139. However, this information is not sufficient to reproduce the data since individual processing steps are not documented in enough detail. Other RIs, such as ICOS and LTER, already provide some provenance information about observation and measurement methods used within the metadata files but are aware that a real tracking tool still needs to be implemented. IAGOS uses the Git software versioning system for code but not for the data themselves. In either case, versioning systems can only be seen as a part of a full provenance infrastructure.

Regarding which information is considered to be important, the answers range from tracking data versioning information to tracking theirgeneration and modification, as well as their usage. This suggests two interpretations about what provenance should comprise:

1. should it enable the community to follow the data ‘back in time’ and to see all the steps that happened from raw data collection, via quality control and aggregation to a useful product or
2. should it enable data providers to track the usage and the users of their data in order to understand its relevance and to improve their services?

Supported by different tools and services, these two roles for provenance may be served within one and the same provenance management system.

As far as domain semantic description of data provenance is concerned, some RIs already use research specific reference tables and thesauri such as EnvThes and SeaDataNet common vocabularies. In this regard there is demand for a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes. Moreover, there is a strong interest among the RIs to learn about provenance in general and, more specifically, to get clear guidance from ENVRIplus about the information provenance should provide. This includes drawing an explicit line between metadata about the “dataset” and explicit provenance information, and whether usage tracking should be part of provenance or not. Another focus is to get support for automated tracking solutions and existing provenance management APIs for their application in the specific e-science environments.

3.2.4.2 Second requirements and use case gathering (2018)

The interviews used in the first round yielded rather moderate and unspecific results. We thus carried out a second gathering round to seek more concrete information from the RIs in order to understand their individual needs regarding the adoption of provenance management systems in their infrastructures. To be effective and to obtain more accurate descriptions without too the end of March 2018, the second round yielded significantly better results, consisting of filled templates returned by nine communities.

More than 20 persons from ten ENVRIplus RIs, including RI representatives and IT experts, joined the initial provenance workshop in Malaga (Spain) in November 2017. During this meeting a ENVRIplus Provenance Working Group was established. A work plan was developed, actions listed and distributed among the involved people[59]. Technically interested RI representatives who regularly followed ongoing discussions, presentations and live demos were asked to provide their requirements and use cases directly, supported by the participating IT experts in an interative process.

The quality of the requirements and use case descriptions, however, was rather unequal. Sometimes RI representatives had no clear understanding about the difference between a use case and a requirement. This had mainly two reasons:

- Whenever a RI was able to send a first draft back to the coordinator (Barbara Magagna) in an early phase of the gathering period an iterative process could take place which helped improving the quality considerably
- Unsurprisingly the level of detail of the information provided via the templates reflected the level of automatization and homogeneity of the IT infrastructures of the RI. RIs like EPOS and IS-ENES, which already implemented at least partial provenance management services, had less difficulties to provide extensive requirements and use cases because they had a very clear idea of what they wanted from provenance in their systems.

[59] https://drive.google.com/open?id=1yIYXdcbvVVELiLILc1KB7KSl9GQLvDLkVjrJtDqc-8A

We think to motivate other RIs to still deliver this input also after the publication of this report and to encourage the RIs involved so far to detail their requirement and use case descriptions as much as possible also in the second half of 2018 so that for end of the year the collection would be completed. The comparison effort (see chapter 0) with use cases of the PPWG's database and with requirements of the Provenance Incubator Group might be helpful in this regard.

For the layout of the requirements and use case template we decided to include some basic elements from Volere Requirements Specification Template[60] as well as from the "The Easy Approach to Requirements Syntax" (EARS)[61]. Aurora Constantin (University of Edinburgh) developed the template as Word Dotx file including definitions and examples for each element as tooltips. An introduction was added to explain the rationale of the gathering and exemplary filled templates by Stephan Kindermann (IS-ENES) as appendix. We found it difficult to work with the MS Wordbased template because of different word versions used by the experts and the limited handling options. This lead to the creation of a Google document[62] with the same elements which could be easily extended (see also Appendix 1).

3.3 Synopsis on ENVRIplus provenance needs

The synopsis provide a compilation of the collected use cases and requirements for each RI. It also compares them among each other (within the different Ris) and with two external resources:

1. Use cases of the RDA PPWG database, which have a persistent URI. For instance 'http://patterns.promsns.org/usecase/62' addresses the UC62 (compare with table references in chapter 3.3.1). For these use cases the PPWG develops specific patterns, which can be valuable recommendations also for ENVRIplus use cases.
2. Requirements of the Provenance Incubator Group, which were derived from 30 own use cases. We will not use the latter use cases for the comparison, but only the requirements, because these are missing in the PPWG database and are general enough to be referred to.

3.3.1 Provenance needs specified for each individual RI

ACTRIS

ACTRIS (Aerosols, Clouds, and Trace gases Research Infrastructure) addresses the scope of integrating state-of-the-art European ground-based stations for long-term observations of aerosols, clouds and short-lived gases[63]. The overall goal of ACTRIS is to provide scientists and other user groups with free and open access to high-quality data about atmospheric aerosols, clouds, and trace gases from coordinated long-term observations, complemented with access to innovative and mature data products, together with tools for quality assurance, data analysis and research. ACTRIS is composed of observing stations, exploratory platforms, instrument calibration centres, and a data centre with three data repositories (also called topic databases): near surface data (EUSAAR), aerosol profiles (EARLINET) and cloud profiles (CLOUDNET).

No specific use case or requirement was provided but a principle interest in being able to generate provenance: The aim is to completely document the execution of the ACTRIS data production workflows. This includes identification of instruments, QC measures on the

[60] http://www.volere.co.uk/template.htm
[61] https://www.iaria.org/conferences2013/filesICCGI13/ICCGI_2013_Tutorial_Terzakis.pdf
[62] https://drive.google.com/open?id=1tdMyY4-RxeddfiKs-lOGLYrmRCfOif5CpO1HHh5Scb0
[63] https://www.actris.eu/

instruments at the atmospheric observatories (type, procedure, by whom, result, documentation), QA measures at central facilities (type, procedure, by whom, result, documentation), QC review process (by whom, data versions produced, issues found), production software (versioned), and all data pre- and final products (versioned). Artifacts used in producing a data pre- or final product are to be referenced in the product's provenance metadata in a standardized way for quantitative accounting for data use.

EMSO

EMSO (the European multidisciplinary seafloor & water column observatory)[64] is a large-scale European Research Infrastructure in the field of environmental sciences for integrating data gathered from a range of ocean observatories. It tries to ensure open access to those data for academic researchers. EMSO is based on a European-scale distributed research infrastructure of seafloor observatories with the basic scientific objective of long-term monitoring, mainly in real-time, of environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere, including natural hazards. It is presently composed of several deep-seafloor observatories, which will be deployed on specific sites around European waters, reaching from the Arctic to the Black Sea passing through the Mediterranean Sea, thus forming a widely distributed pan-European infrastructure. A goal of EMSO is to harmonise data curation and access, while averting the tendency for individual institutions to revert to idiosyncratic working practices after any particular harmonisation project has finished.

No specific use case or requirement was provided but the general aim to track provenance data which should be used within the EMSO data portal metadata catalogue.

EISCAT-3D

EISCAT-3D[65] is a research infrastructure that will use a new generation of phased array radars to study the Earth's middle atmosphere, ionospheric incoherent scatter and objects in space, contributing to near-Earth space environment research. It aims at establishing a system of distributed phased array radars. The system will enable comprehensive three-dimensional observations of ionospheric parameters and atmospheric dynamics above Northern FennoScandinavia, which is an important location for research on coupling between space and the polar atmosphere. EISCAT-3D will produce about 2 petabytes of data each year and aims at using standard systems for data storage and cataloguing, user authentication and identification and citation of datasets.

EISCAT -3D provided use cases and requirements descriptions[66].

UC Nr	R Nr	Name	Rel UC	Rel P_UC
E3D.U1	E3D.R1	VRE support for user-driven analysis		UC47, UC57

Table 1: EISCAT-3D use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
E3D.R1	E3D.U1	Software as metadata	ICOS.R11, LTER.R3.3	

[64] http://www.emso-eu.org/
[65] https://www.eiscat.se/
[66] https://drive.google.com/open?id=1fSk5MI_LVhYJl6qaZWfj4GNOKg90MpqvIE-XPe2uEC4

E3D.R2	Parameter and settings	IS-ENES.R1, ICOS.R11, LTER.R3.2, ANAEE.R3	C-PROC-UR3
E3D.R3	Quality control and flagging	ICOS.R7	C-JUST-UR2

Table 2: EISCAT-3D requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

IS-ENES

The European Network for Earth System Modelling (IS-ENES2) is the second phase of the I3 infrastructure project for the European Network for Earth System Modelling (ENES). The third phase is proposed and submitted to the European Union. ENES gathers the community working on climate modelling. IS-ENES runs a distributed, federated data infrastructure based on a few (3-4) main data centres and various associated smaller ones and coordinates (and operates) the European contribution to the worldwide ESGF infrastructure. IS-ENES encompasses climate models and their environment tools, model data and the interface of the climate modelling community with high-performance computing, in particular the European RI PRACE.

The IS-ENES infrastructure tries to capture the whole data life cycle from model based data generation to data distribution and processing to data archival and data citation.

IS-ENES provided use cases and requirements descriptions[67] along this data life cycle.

UC Nr	R Nr	Name	Rel UC	Rel P_UC
IS-ENES.U1	IS-ENES.R1	Scientific data provenance	ICOS.U5, LTER.U3, EPOS.U5	UC54
IS-ENES.U2	IS-ENES.R2	Provenance for derived data products		UC49
IS-ENES.U3	IS-ENES.R3	Data ingest provenance	ICOS.U4	UC34
IS-ENES.U4	IS-ENES.R4	Data versioning and errata tracking		UC43
IS-ENES.U5		Long term data archival		UC41

Table 3: ICOS use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
IS-ENES.R1	IS-ENES.U1	Scientific provenance for model generated datasets	E3D.R2, ICOS.R11, LTER.R3.2, ANAEE.R3	C-PROC-UR3,
IS-ENES.R2	IS-ENES.U2	Provenance for derived data products	ANAEE.R10	
IS-ENES.R3	IS-ENES.U3	Data ingest provenance		U-Inter-UR2
IS-ENES.R4	IS-ENES.U4	Data versioning and errata tracking		U-Debug-UR1

[67] https://docs.google.com/document/d/1Q7b5SBB6zgSidu4kBH-c1jv1vAjp5dqGA_9WPgQ30Cw/edit?usp=sharing

Table 4: IS-ENES requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

ANAEE

AnaEE (Analysis and Experimentation on Ecosystems)[68] focuses on providing innovative and integrated experimentation services for ecosystem research. It will strongly support scientists in their analysis, assessment and forecasting of the impact of climate and other global changes on the services that ecosystems provide to society. The gathering of information in a common portal should help with this.

AnaEE provided only requirements descriptions[69].

R NR	UC Nr	Name	Rel R	Rel PI_R
ANAEE.R1		Provenance for experimental facilities		
ANAEE.R2		Provenance for the variable/trait observed	EPOS.R2	U-Under-UR1
ANAEE.R3		Provenance for the experimental design	IS-ENES.R1, E3D.R2, LTER.R3.2, ICOS.R11	C-Proc-UR3 U-Under-UR1
ANAEE.R4		Provenance for the spatial and temporal context		
ANAEE.R5		Provenance for the data acquisition tool	ICOS.R.3, LTER.R1.5	
ANAEE.R6		Provenance for actors	ICOS.R2, LTER.R1.1	C-Attr-UR1
ANAEE.R7		Provenance for data curation process	EPOS.R3	
ANAEE.R8		Provenance for data annotation process	LTER.R2.2	U-Under-UR1
ANAEE.R9		Provenance for the metadata generation process		U-Under-UR2
ANAEE.R10		Provenance for the dataset generation/curation/publication	IS-ENES.R2	

Table 5: ANAEE requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

LTER-EUROPE

Long-Term Ecosystem Research (LTER) is an essential component of worldwide efforts to better understand ecosystems. LTER Europe[70] aims at providing information on ecosystem functioning and processes as well as related drivers and pressures for a whole ecosystem (e.g., a watershed). This information is very diverse in its technical formats (sensor Information, aerial photographs, field recordings, pictures, etc.). The purpose of the RI is to focus on harmonised methodologies and data products. Due to the fragmented character of LTER Europe, harmonised data documentation, real-time availability of data as well as harmonisation of data and data flows are the overarching goals for the forthcoming years. Currently, LTER Europe is developing a Data Integration Portal (DIP, e.g. including a time series viewer) and is working on the integration of

[68] https://www.anaee.com/

[69] https://drive.google.com/open?id=1z1pKoZTTIA94PV38mbN--mxSMISvAJQT853pbMg_lFI

[70] http://www.lter-europe.net/

common data repositories into their workflow system (including metadata documentation with LTER Europe DEIMS).

Due to its holistic approach to conceive the ecosystem, Long Term Ecosystem Research is characterized by the heterogeneity of existing approaches. LTER sites around the globe apply a wide variety of data acquisition methods, use different means to store, exchange and process data. As far as the collection of provenance information is concerned, three main use cases[71] have been derived in this regard, dealing with provenance for data acquisition, data aggregation of heterogeneous content, as well as for script based data processing.

UC Nr	R Nr	Name	Rel UC	Rel P_UC
LTER.U1		Provenance for data acquisition		
LTER.U1.1	LTER.R1.1, LTER.R1.2	Non-automated data collection-observation		
LTER.U1.2	LTER.R1.1, LTER.R1.2, LTER.R1.3, LTER.R1.4	Non-automated data collection-physical samples	ICOS.U2	
LTER.U1.3	LTER.R1.5	Automated data collection via sensors		
LTER.U2	LTER.R2.1, LTER.R2.2	Track lineage for heterogeneous data sources via scientific publications		UC21
LTER.U3	LTER.R3.1, LTER.R3.2, LTER.R3.3	Track lineage for ad-hoc workflows combining scientific scripts and third party software	IS-ENES.U1, ICOS.U5, EPOS.U5	UC54

Table 6: LTER use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
LTER.R1.1	LTER.U1.1	Registry for persons	ICOS.R2, ANAEE.R6	C-Attr-UR1
LTER.R1.2	LTER.U1.1	Trace Provenance in Excel or other spreadsheets		
LTER.R1.3	LTER.U1.2	Registry for lab equipment	ICOS.R2	
LTER.R1.4	LTER.U1.2	Registry and representation for manual workflows	ICOS.R6	
LTER.R1.5	LTER.U1.3	Registry for sensors and measurement devices	ICOS.R3, ANAEE.R5	
LTER.R2.1	LTER.U2	Registry for publications		
LTER.R2.2	LTER.U2	Semantic annotation of method descriptions in publications	ANAEE.R8	
LTER.R3.1	LTER.U3	Track data processing steps within scientific scripts	ICOS.R10	U-Inter-UR2
LTER.R3.2	LTER.U3	Tracking model runs/used parameters	IS-ENES.R1, E3D.R2, ICOS.R11, E3D.R1	C-Proc-UR3
LTER.R3.3	LTER.U3	Software provenance/Software registry		

[71] https://drive.google.com/open?id=1VJPnJJGiOZNYXIX8Ao6g3LSXHC6HyTuu9Vplsvtu0Gc

Table 7: LTER requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

ICOS

ICOS ERIC is a highly distributed RI, with observation station networks located in twelve member states and the key organizational components hosted by nine countries.[72] To not only support a high degree of data and metadata interoperability within ICOS itself, but to ensure and guarantee (re-)usability of all ICOS data products by a wide range of end user communities, it is absolutely critical to collect, catalogue and disseminate correct and comprehensive metadata, including provenance.

Figure 23 shows a schematic of the data flow in ICOS, with the letters A-I indicating distinct entities or actors, which should collect provenance information of different kinds:

A. Measurement stations
B. The Atmospheric Thematic Centre and the Central Analytical Laboratories
C. The Ecosystem Thematic Centre
D. The Oceanic Thematic Centre
E. The Carbon Portal including the ICOS data (and metadata) repository
F. End users of ICOS data products
G. Producers of elaborated data products
H. External metadata services (portals)
I. HTC and HPC computing processing facilities

ICOS provided two well documented examples ("Documenting condition changes at an observation station" and "Preparing for dissemination of end user-produced 'elaborated' data products"), as well as a number of use cases and requirements[73].

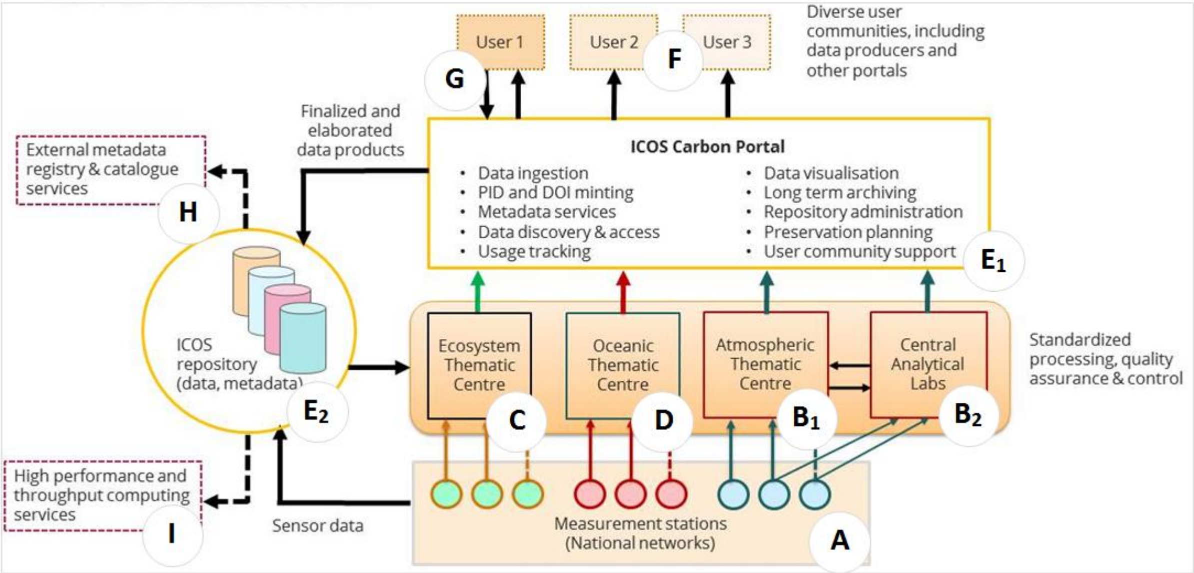

User 1
User 2
User 3
Diverse user communities, including data producers and other portals
G
F
Finalized and elaborated data products
External metadata registry & catalogue services
H
ICOS Carbon Portal
Data ingestion
PID and DOI minting
Metadata services
Data discovery & access
Usage tracking
Data visualisation
Long term archiving
Repository administration
Preservation planning
User community support
E1
ICOS repository (data, metadata)
E2
Ecosystem Thematic Centre
Oceanic Thematic Centre
Atmospheric Thematic Centre
Central Analytical Labs
C
D
B1
B2
Standardized processing, quality assurance & control
High performance and throughput computing services
I
Sensor data
Measurement stations (National networks)
A

Figure 23: A schematic illustration of the data flow in ICOS, from measurement stations, via processing at Thematic Centres, to curation and dissemination at the Carbon Portal.

[72] https://www.icos-cp.eu/
[73] https://drive.google.com/open?id=1tqTRr2jh5sknwK-3afJAWq-mlxVRloASmYecZfhdY8g

UC Nr	R Nr	Name	Rel UC	Rel P_UC
ICOS.U1	ICOS.R1, ICOS.R2, ICOS.R3, ICOS.R4	Change of Measurement Station Information		UC57
ICOS.U2	ICOS.R5, ICOS.R6	Handling of physical samples	LTER.U1.2	
ICOS.U3	ICOS.R7	Eddy Covariance data algorithms	IS-ENES.U1, LTER.U3, EPOS.U5	UC54
ICOS.U4	ICOS.R8, ICOS.R9	Carbon Portal register	IS-ENES.U3, EMBRC.U1	UC2, UC24
ICOS.U5	ICOS.R10, ICOS.R11	Elaborated data production specifics	IS-ENES.U1, LTER.U3, EPOS.U5	UC54

Table 8: ICOS use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
ICOS.R1	ICOS.U1	Change in site environment		
ICOS.R2	ICOS.U1	Registry of personnel	ANAEE.R6, LTER.R1.1	C-Attr-UR1
ICOS.R3	ICOS.U1	Registry of sensors & instrumentation	ANAEE.R5, LTER.R1.5	
ICOS.R4	ICOS.U1	Event based recording of changes		C-Vers-UR1/1b/2b/3b
ICOS.R5	ICOS.U2	Identifier for physical samples	LTER.R1.3	
ICOS.R6	ICOS.U2	Manual provenance	LTER.R1.4	
ICOS.R7	ICOS.U3	Quality control and flagging	E3D.R3	C-JUST-UR2
ICOS.R8	ICOS.U4	Track usage of data objects	EMBRC.R1	
ICOS.R9	ICOS.U4	Harvest bibliometric databases		
ICOS.R10	ICOS.U5	Capturing input data lineage	LTER.R3.1	U-Inter-UR2
ICOS.R11	ICOS.U5	Capturing used algorithms, models and software	E3D.R2, IS-ENES.R1, LTER.R3.2	C-Proc-UR3

Table 9: ICOS requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

EMBRC

EMBRC (European Marine Biological Resource Centre)[74] is a distributed European RI which is set up to become the major RI for marine biological research, covering everything from basic biology, marine model organisms, biomedical applications, biotechnological applications,

[74] http://www.embrc.eu/

environmental data, ecology, etc. Having successfully completed a 3-year preparatory phase (2011-2014) and an implementation phase (2014-2016), and operation has already started 2016-2017. The main purpose of EMBRC is to promote marine biological science and the application of marine experimental models in mainstream research by providing the facilities (lab space), equipment (e.g., electron microscopes, real time PCR machines, crystallography, lab equipment, equipment for accessing the environments such as research vessels, scientific divers, ROVs, etc.), expertise and biological resources that are necessary for carrying out biological research In what concerns data, the role of EMBRC is to generate and make it available. It does not usually do any analysis of those data, unless it is contracted to do so. Data is usually generated through sensors in site in the sea or samples that are collected and then measured in the lab.

EMBRC provided one use case with one requirement[75].

UC Nr	R Nr	Name	Rel UC	Rel P_UC
EMBRC.U1	EMBRC.R1	Provenance information for tracking disparate data use	ICOS.U4	UC35, UC24, UC23, UC2

Table 10: EMBRC Use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
EMBRC.R1	EMBRC.U1	EMBRC Data Aggregation	ICOS.R8	U-Inter-UR1

Table 11: EMBRC requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

EPOS

EPOS is a long-term plan for the integration of Research Infrastructures for Solid Earth Science in Europe. Its main aim is to integrate communities to make scientific discovery in the domain of solid earth science.[76] EPOS integrates the existing (and future) advanced European facilities into a single, distributed, sustainable infrastructure (EPOS Core Services) taking full advantage of new e-science opportunities. EPOS will allow the Earth Science community to make a significant step forward by developing new concepts and tools for accurate, durable, and sustainable answers to societal questions concerning geo-hazards and those geodynamic phenomena (including geo-resources) relevant to the environment and human welfare.

EPOS provided use cases and requirements descriptions[77].

UC Nr	R Nr	Name	Rel UC	Rel P_UC
EPOS.U1	EPOS.R2	Discovery of experiments and data		UC54, UC49
EPOS.U2		Data and dependencies navigation		UC22
EPOS.U3	EPOS.R4	Monitoring		
EPOS.U4		Preview and staging		
EPOS.U5	EPOS.R5	Reproducibility	IS-ENES.U1, LTER.U3,	UC54

[75] https://drive.google.com/open?id=16LO_Ef04QaLkr5xyCS67brgMVePA9XNoFBsoGOIvP4w
[76] https://www.epos-ip.org/
[77] https://drive.google.com/open?id=1uvkB6oOijWFgOxFv_jO4u5rD_6tjzuxde3H08Slb030

		ICOS.U5	
EPOS.U6	Configuration of interdependent tasks (data reuse)		
EPOS.U7	Analysis of collaborative interactions and data reuse		UC2, UC35, UC23
EPOS.U8	Selective transfer		UC21
EPOS.U9	Selective generation of traces		

Table 12: EPOS use cases (UC Nr) with related requirements (R Nr), related other use cases within ENVRIplus (Rel UC) and related use cases of the PPWG (Rel P_UC)

R NR	UC Nr	Name	Rel R	Rel PI_R
EPOS.R1	EPOS.U1-U9	Results Validation -1 (agent)		C-Attr-UR2
EPOS.R2	EPOS.U1	Results Validation -2 (domain specific)	ANAEE-R2	U-Under-UR1
EPOS.R3	EPOS.U4-U5	Results Validation -3 (granularity)		C-Proc-UR5
EPOS.R4	EPOS.U3	Methods Validation (software dependencies)		
EPOS.R5	EPOS.U5	Worfklow's execution configurations	IS-ENES.R1, E3D.R2, LTER.R3.2, ICOS.R11	C-Proc-UR3
EPOS.R6	EPOS.U1-U9	Storage and access (query)		M-Acc-UR1

Table 13: EPOS requirements (R Nr) with related use caes (UC Nr), related other requirements within ENVRIplus (Rel R) and related requirements of the PI (Rel PI_R)

3.3.2 Synopsis on use cases

From a system level point of view, one of the most prominent differences between ENVRIplus RIs is their varying level of automation. Some RIs are built on fully automated sensor networks where human intervention is mostly limited to monitoring, interpretation and maintenance tasks. Other RIs in turn include significantly more manual steps, taking place in data acquisition, exchange, or even processing. This diversity is clearly reflected in the use cases reported by the different RIs.

In more automated settings, e.g. EPOS, the reported use cases are often targeted towards clearly defined user needs and system features to address them, such as “Discovery of experiments”, “Navigation through data and dependencies”, etc.

In less automated settings, reported Use Cases appear more targeted towards the different aspects of provenance collection itself, e.g. which different scenarios exist where such information shall be stored, and less on subsequent applications of such data. Stated use cases include tracking lineage for script based workflows, provenance for automated and non-automated data acquisition such as human observation and physical sample based data collection, and provenance for data publishing and reuse.

3.3.3 ENVRIplus use cases to be included as new Use Cases in the PPD

The following use cases are proposed to be included as new Use Cases in the Provenance Patterns Database, as no related could be found:

UC Nr	Name
LTER.U1.1	Non-automated data collection-observation
LTER.U1.2, ICOS.U2	Non-automated data collection-physical samples
LTER.U2	Track lineage for heterogeneous data sources via publications
LTER.U3	Track lineage for ad-hoc workflows combining scientific scripts and third party software
EPOS.U3	Monitoring
EPOS.U4	Preview and staging
EPOS.U6	Configuration of interdependent tasks (data reuse)
EPOS.U9	Selectice generation of traces

Table 14: Proposed new PPD-UC from ENVRIplus

3.3.4 Synopsis on requirements

On the level of expressed requirements, the different RIs converge more. A central element are various types of registries, since tracking provenance for processes with different agents and entities usually requires unique identifiers for each involved instance. Registries for persons, data objects, measurement devices etc. can thus be almost considered a prerequisite for meaningful provenance approaches.

Other commonly expressed requirements include storing provenance for model parameters or workflow configurations, including domain specific metadata from controlled vocabularies in the provenance tracks and tracking of data use/citations.

4 BEST PRACTICES AND IMPLEMENTATIONS

4.1 RDA Provenance Patterns

The motivation of the RDA Provenance Patterns WG is the conviction that some ways of doing things in provenance are better than others. Apart from collecting provenance use cases the second activity area is about generating provenance design patterns for provenance task such as representation, transmission, use etc. Both use cases and patterns are stored in an online database[78]. The system supports linking between related items, e.g. broader and narrower (specialised) Use Cases and broader and narrower Patterns and also between Use Cases and Patterns that address them.

Provenance Patterns are best practice guides for accomplishing certain, provenance-related, tasks. A pattern is an abstract generalization of a solution for a use case collected in the database. More generally, a solution pattern represents an interpretive response on a use case, highlighting key decisions and recommendations[79].

A Pattern is composed of the following structural elements: Contributor, broader Pattern, related Use Case, introduction, prerequisite, implementation, example, summary.

[78] http://patterns.promsns.org

[79] https://rd-alliance.org/group/provenance-patterns-wg/case-statement/working-group-provenance-patterns-case-statement

The Pattern choice is justified on the basis of stakeholder communities with whom one wishes to cooperate. Computational limitations or costs can also influence the decision to adopt or reject a pattern. In the Pattern's description sometimes current document practices are marked as Antipatterns, followed by explanations, why these approaches are not considered as best practices.

The recommendations are matched to the granularity of the underlying use cases. The Patterns don't match the abstraction level of the use case, but aim for a level that highlights key modelling and encoding decisions.

So far (April 2018) five fully described Provenance Patterns are provided in the database, a few others are not yet completed:

PP Nr	Name
PP18	A basic data processing model
PP12	Associating metadata in documents with graph provenance
PP13	Describing activities' actions
PP63	Agent Role Patterns
PP25/PP26	Associating ISO19115-1 items with a provenance query service

Table 15: Provenance Patterns

4.2 EPOS

EPOS consists of a central integrated core services (ICS) facility which provides a gateway to ten thematic core services (TCS) facilities. These thematic core services bring together—at a European scale—multiple European, national and institutional research infrastructures (over 250 in total). The TCSs vary greatly in the number of assets they provide access to and their complexity, and consequently their requirements for provenance vary as well. Although all TCSs have concerns about and interest in provenance, very few have an implemented provenance system. Some have a basic tracking system for activities related to datasets, some others cover computational workflows and offer repositories and Web APIs for its management.

4.2.1 EPOS data lifecycle

For all of the EPOS community, provenance is bound intimately with EPOS catalogue services (including both local TCS catalogues and the central ICS catalogue) and with curation. It is also linked intimately with checkpointing, recovery and accounting. Finally, it is required for provenance to record authorisations based on authentication, and thus has to interface with authorisation and authentication services. The EPOS data lifecycle for TCSs (including their interactions with ICS) is shown below, overlaid with the provenance cycle. This illustrates the requirements for provenance reporting as seen in EPOS - as something integrated with the metadata catalogue, curation and processing control, operating throughout the lifecycle.

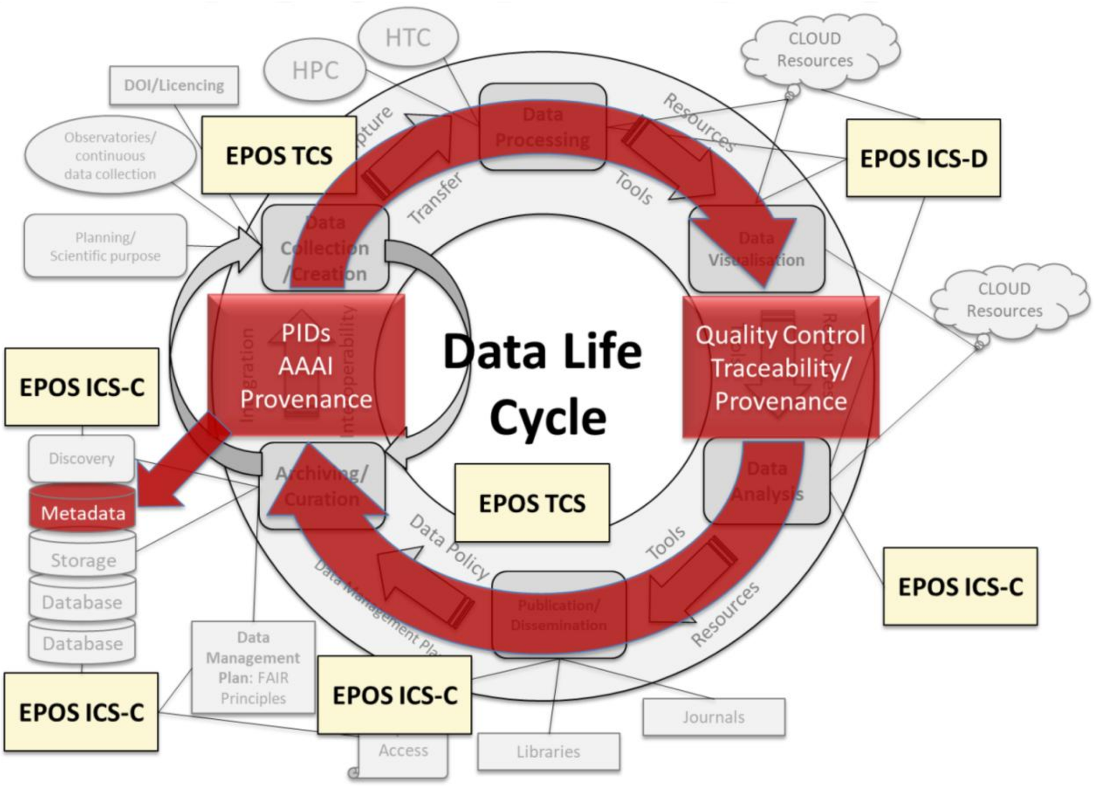


Figure 24: EPOS TCS lifecycle, overlaid with provenance/QC cycle (Atakan et al., EPOS WP6)

More sophisticated provenance tracking has been addressed in the context of the FP7 VERCE project however, which contributes specifically to the seismology TCS, but has been developed using generic technologies and standards (i.e. PROV), making the work applicable to EPOS (and other research infrastructures) more generally.

4.2.2 VERCE

The VERCE project within the context of the seismology theme has been experimenting with the use of PROV in workflows executed through a VRE (Virtual Research Environment). The VERCE project has pioneered e-infrastructure to support researchers using established earthquake simulation codes on high-performance computers in conjunction with the misfit analysis of the simulation results with observational data obtained from multiple sources [Atkinson et al., 2015]. This is accessed and organised via the VERCE science gateway, which makes it convenient for seismologists to use these resources from any location via the Internet. Their data handling is made flexible and scalable by two Python libraries, ObsPy and dispel4py, and by data services delivered by institutional federated archives (i.e. FDSN). In this context, a provenance management system, S-ProvFlow (described in more detail below and particularly in [Spinuso, 2018b]), enables monitoring and validation of each computation performed through the portal according to its underlying model S-PROV (compare chapter 2.4.8). This system together with specific provenance supporting tools accomodates all EPOS use cases and requirements mentioned in the chapter 3.3.1. In the following we try to specify which component supports which use case scenario and related requirement.

S-ProvFlow offers user interfaces in support of these actions, allowing researchers to manage their results, enabling rapid exploration of results and of the relationships between data. Moreover, it allows through its Web API for the semi-automatic configuration of interdependent workflows' parameters and inputs (see Figure 25), supporting EPOS.U5, EPOS.U6 and EPOS.R5, as described in chapter 3.3.1. Spinuso [2018b] introduced also the concept of Active Provenance for assisted usability for data-intensive computations. It allows contextualisation and selectivity of the scope and nature of the lineage which users consider relevant by tuning the precision of provenance capture, supporting EPOS.U8 and EPOS.U9.

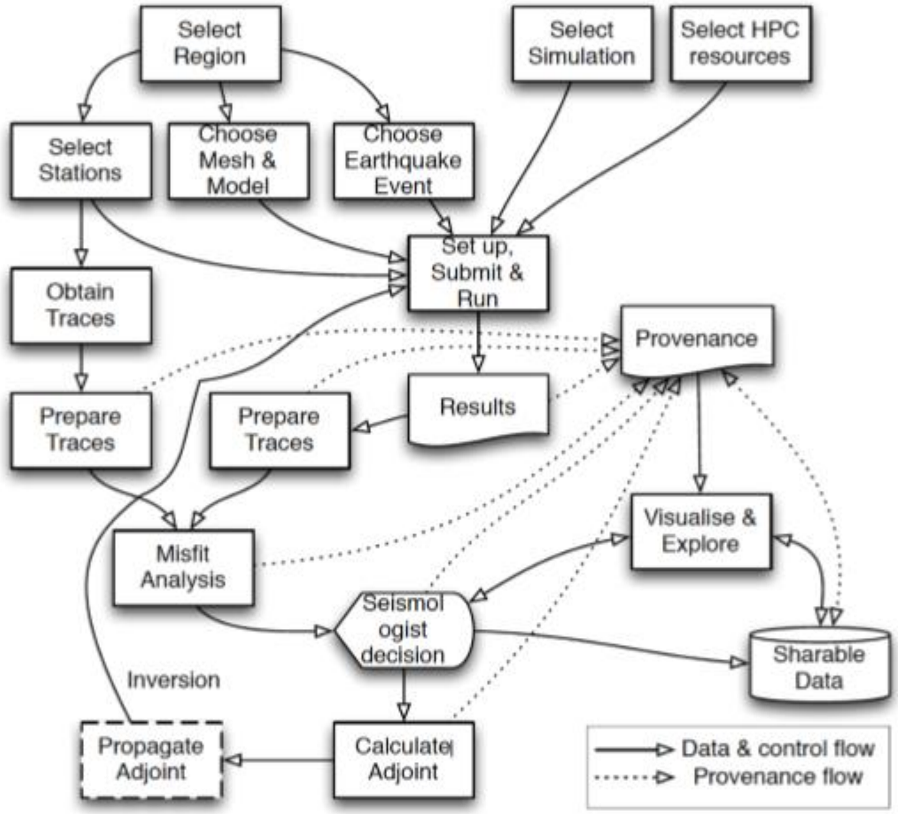


Figure 25: Central role of Provenance information in the forward wave propagation, misfit analysis and inversion [Spinuso, 2018b]

Figure 26 shows the General Computational Pattern adopted in VERCE to represent invocation of stateful processes, extracted from the S-PROV model. These typically belong to workflow systems and access data, such as intermediate results or external resources, beyond their direct inputs.

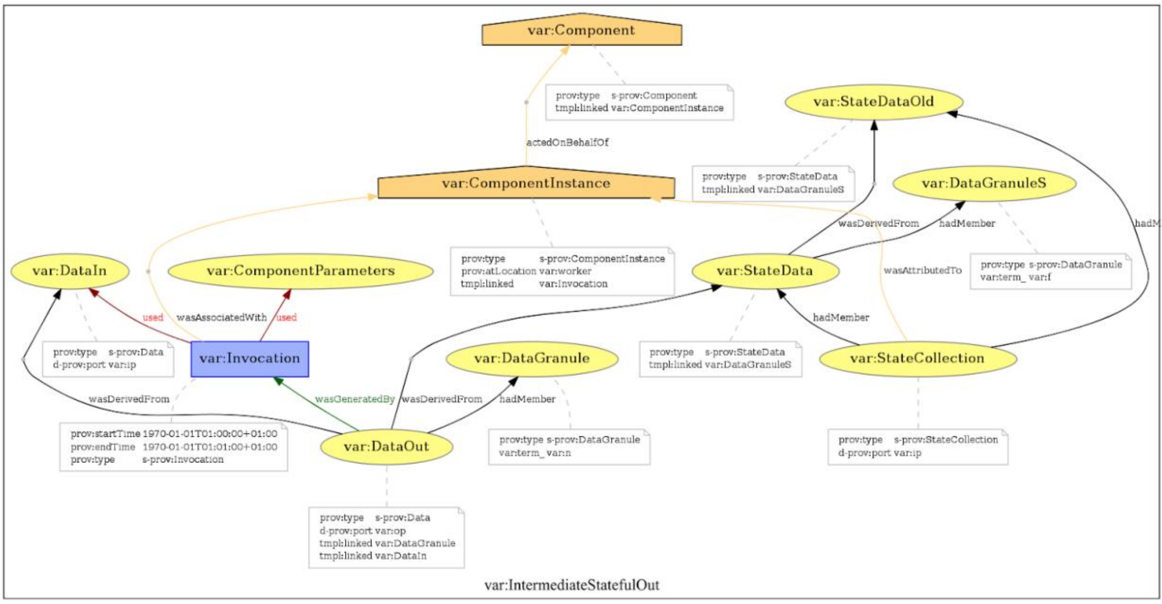

var:Component
prov:type s-prov:Component
tmpl:linked var:ComponentInstance
actedOnBehalfOf
var:StateDataOld
var:ComponentInstance
prov:type s-prov:StateData
tmpl:linked var:DataGranuleS
var:DataGranuleS
wasDerivedFrom
hadMember
hadMember
var:DataIn
var:ComponentParameters
prov:type s-prov:ComponentInstance
prov:atLocation var:worker
tmpl:linked var:Invocation
var:StateData
wasAttributedTo
prov:type s-prov:DataGranule
var:term_ var:f
used
wasAssociatedWith
used
hadMember
prov:type s-prov:Data
d-prov:port var:ip
var:Invocation
var:DataGranule
prov:type s-prov:StateData
tmpl:linked var:DataGranuleS
var:StateCollection
wasDerivedFrom
wasGeneratedBy
wasDerivedFrom
hadMember
prov:startTime 1970-01-01T01:00:00+01:00
prov:endTime 1970-01-01T01:01:00+01:00
prov:type s-prov:Invocation
var:DataOut
prov:type s-prov:DataGranule
var:term_ var:n
prov:type s-prov:StateCollection
d-prov:port var:ip
prov:type s-prov:Data
d-prov:port var:op
tmpl:linked var:DataGranule
tmpl:linked var:DataIn
var:IntermediateStatefulOut

Figure 26: VERCE General Computational Pattern for stateful process invocation

The template represents the invocation of a process (the Invocation activity in blue) from a software agent (the ComponentInstance agent in orange) acting on behalf of an abstract workflow component (Component in orange), as represented by the S-PROV model. Updates to the StateCollection of a ComponentInstance are traced across Invocations (StateData was derived from StateDataOld). The Data entity is a general container which may contain additional DataGranules.

A Data entity is characterised by a set of baseline metadata properties: size, format, annotation, location, write-count, immediate-access, first-known-destination, etc. The last two properties listed support the use case of controlled runtime transfers of a data item to a known external resource or system for post-processing or visualisation.

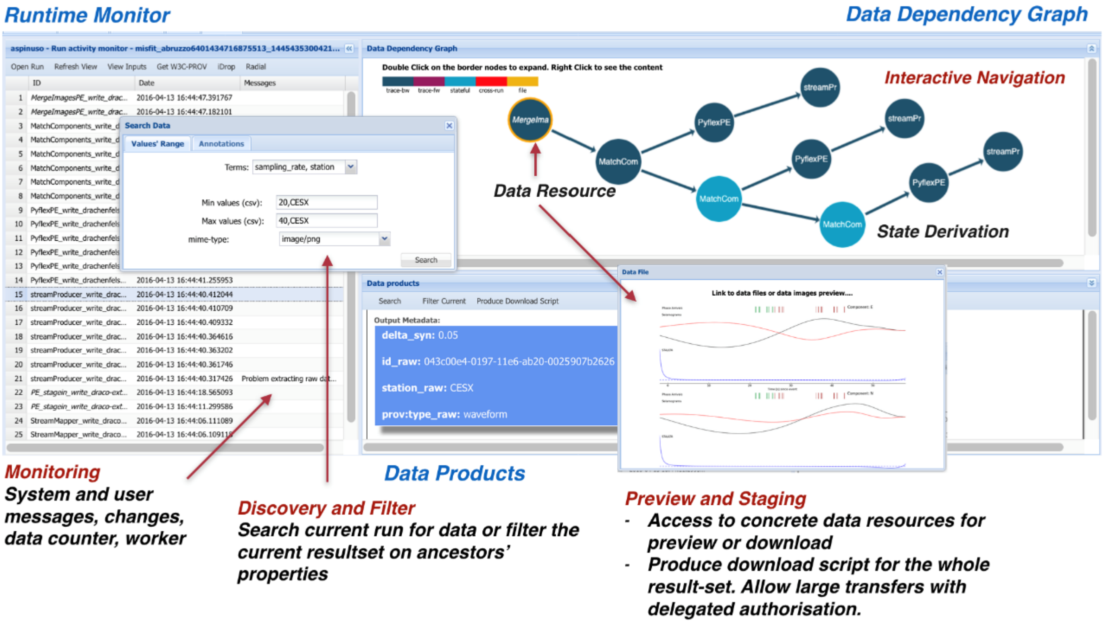
A DataGranule is a member of a Data entity and is specific to the application domain running the workflow (allowing for scenarios where a workflow can serve multiple domains, thus decoupling the internal data structures from their semantics within specific contexts). It is described by a set of properties belonging to a community vocabulary or introduced by the user. This representation supports this requirement EPOS.R2 (compare chapter 3.3.1, EPOS).

4.2.3 S-ProvFlow

The S-ProvFlow system offers a set of components that support provenance acquisition and exploration. It includes a database, a Web service layer and two complementary interactive tools.

One of the tools, the Monitoring and Validation Visualiser (MVV), Figure 27, assists the users in the fine-grain interpretation of the provenance records in order to understand dependencies; it allows them to select and configure viewpoints by specifiable searches over domain metadata value-ranges, previews, navigation of data dependency graph, within and across runs, data download and staging. It offers detailed runtime diagnostics also differentiating between stateless and stateful operations. It enables the selective export of traces in PROV-XML and RDF, which can cover full runs or single traces associated with a specific data entity. This tool supports EPOS.U2 – Data and Dependencies Navigation, EPOS.U3 – Monitoring, EPOS.U4 – Preview and Staging, EPOS.R3 and EPOS.R4 requirements (see chapter 3.3.1, EPOS).

A graphical tool, the Bulk Dependency Visualiser (BDV), Figure 28, offers broader perspectives on computational characteristics as well as collaborative behaviour via customisable radial diagrams. It adopts hierarchical edge bundle techniques and configurable grouping. It allows its users to dynamically adjust viewing and clustering controls to uncover aspects of the distribution of the processing for large single runs, as well as data-reuse between different workflows' executions and users thus supporting EPOS.U7.



Runtime Monitor

Data Dependency Graph

aspinuso - Run activity monitor - misfit_abruzzo6401434716875513_1445435300421...

	ID	Date	Messages
1	*MergeImagesPE_write_drac...*	2016-04-13 16:44:47.391767	
2	*MergeImagesPE_write_drac...*	2016-04-13 16:44:47.182101	
3	MatchComponents_write_d		
4	MatchComponents_write_d		
5	MatchComponents_write_d		
6	MatchComponents_write_d		
7	MatchComponents_write_d		
8	MatchComponents_write_d		
9	PyflexPE_write_drachenfels		
10	PyflexPE_write_drachenfels		
11	PyflexPE_write_drachenfels		
12	PyflexPE_write_drachenfels		
13	PyflexPE_write_drachenfels		
14	PyflexPE_write_drachenfels...	2016-04-13 16:44:41.255953	
15	streamProducer_write_drac...	2016-04-13 16:44:40.412044	
16	streamProducer_write_drac...	2016-04-13 16:44:40.410709	
17	streamProducer_write_drac...	2016-04-13 16:44:40.409332	
18	streamProducer_write_drac...	2016-04-13 16:44:40.364616	
19	streamProducer_write_drac...	2016-04-13 16:44:40.363202	
20	streamProducer_write_drac...	2016-04-13 16:44:40.361746	
21	streamProducer_write_drac...	2016-04-13 16:44:40.317426	Problem extracting raw dat...
22	*PE_stagein_write_draco-ext...*	2016-04-13 16:44:18.565093	
23	*PE_stagein_write_draco-ext...*	2016-04-13 16:44:11.299586	
24	StreamMapper_write_draco...	2016-04-13 16:44:06.111089	
25	StreamMapper_write_draco...	2016-04-13 16:44:06.109118	

Search Data

Terms: sampling_rate, station

Min values (csv): 20,CESX

Max values (csv): 40,CESX

mime-type: image/png

Data Dependency Graph

Double Click on the border nodes to expand. Right Click to see the content

Interactive Navigation

Data Resource

State Derivation

Data products

Output Metadata:

delta_syn: 0.05

id_raw: 043c00e4-0197-11e6-ab20-0025907b2626

station_raw: CESX

prov:type_raw: waveform

Data File

Link to data files or data images preview....

Monitoring

System and user messages, changes, data counter, worker

Discovery and Filter

Search current run for data or filter the current resultset on ancestors' properties

Data Products

Preview and Staging

- Access to concrete data resources for preview or download
- Produce download script for the whole result-set. Allow large transfers with delegated authorisation.

Figure 27: MVV: Navigable graphical representation and combined access to data products and metadata. Snapshot showing the provenance analysis of a Misfit Workflow

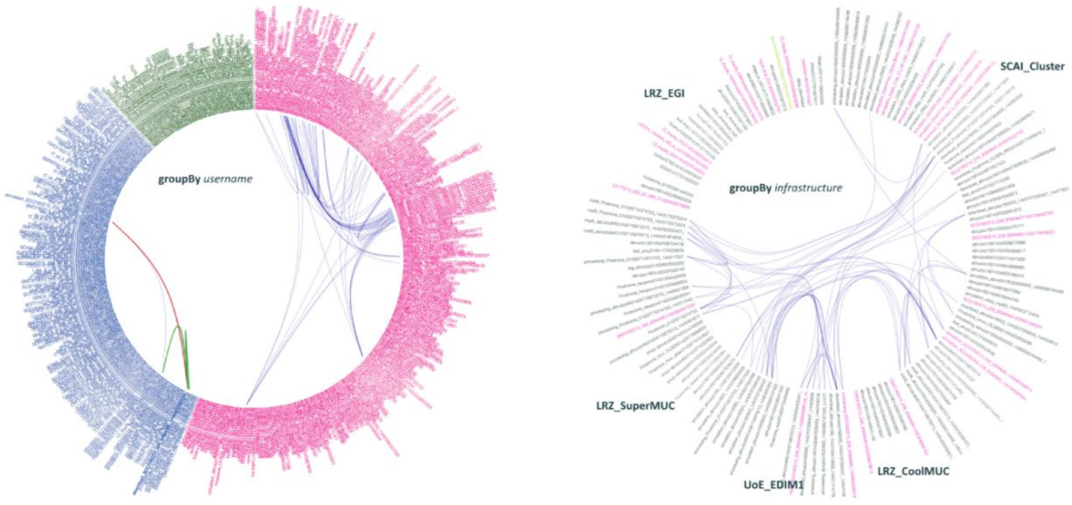


Figure 28: BDV: collaborative interactions among users workflows and infrastructures. The diagrams are obtained by searching for runs that involved data that present metadata within specific value-ranges and applying two different grouping rules (a) by user-name

The S-ProvFlow system exposes a RESTFul web API which offers high-level services on top of the storage backend implemented in MongoDB. The API methods provide a generic class of provenance methods: provenance ingestion (ping-back approach for runtime updates), monitoring, detailed metadata and lineage exploration, summarisation, integrated data and experiments discovery, thus it supports EPOS. UC1 (compare 3.3.1, EPOS). Discovery is supported by using the provenance metadata terms which are part of the underlying model, as well as domain specific terms that may be added to the collection by different communities and for different experiments. These can belong to controlled vocabularies or being introduced ad-hoc for experimental purposes (EPOS.R1). The API serves different workspaces of the VERCE gateway and and feeds those components of the user interface that allow the reload of the configurations of past experiments or that provide to the researcher indicators and semi-automated mechanisms to discover and access those results that can be combined to initiate misfit analysis (EPOS.R6).

The S-ProvFlow system will be further developed within EPOS and in the context of the H2020 DARE project. DARE identifies a number of use cases to which the S-ProvFlow system will be applied, including estimating mean strong ground motions, models of strong ground motions and rapid characterisation of seismic sources.

4.3 IS-ENES

4.3.1 Scientific background and application domain

In the climate sciences, there are two rough categories when it comes to data:

3. **Modelling data** is born digital, generated by running complex software code (climate models) on HPC systems.
4. **Observational data** is measured by sensors, for example through remote sensing (satellite data).

In both cases, the variables contained in the data are possibly of wide range, but in principle related to the same conceptual entities, covering physical areas such as atmosphere, oceans and other water bodies, sea ice, land usage and so on. Variables may therefore include, for example, air temperature, wind speed, ocean surface temperature, salinity, vegetation coverage and so on. The use case imagined for data typing and broker usage is **to enhance combined modelling and observational data processing.**

IS-ENES concentrates on the modelling data domain yet integrating observational data to evaluate modelling results and to configure and tune the climate models. Scientific groups and research institutions around the globe develop individual climate models, which are run on their respective HPC systems. However, there is no perfect climate model, and all of them model the physical world in different ways. To assess the quality of climate models, a large internationally coordinated exercise is therefore needed: Running the various models with same input and boundary conditions, producing data conforming to agreed standards and conventions that can then be analyzed and compared to assess the differences between models or to generate aggregated “ensemble” data products (basic statistics). These exercises are called the Model Intercomparison Projects (MIPs). The largest of these exercises driving the infrastructural developments is called Coupled Model Intercomparison Project (CMIP)[80].

CMIP is in essence a cyclic activity, historically divided into phases. Each phase runs for several years. The previous phase, now finished, was CMIP5; the current phase is called CMIP6. The first model runs producing data are estimated to start in 2017, with the whole phase’s operations

[80] https://www.wcrp-climate.org/wgcm-cmip

stretching over roughly the next 3-4 years. Organizationally, CMIP is coordinated under the umbrella of the World Climate Research Programme (WCRP).

While intercomparison of model outputs is one primary driver of CMIP, its output represents a highly valuable collection of data produced by state-of-the-art climate models, enabling further downstream usage. Most importantly, the Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC) are based on publications that are written on CMIP data, and the schedule of CMIP and the IPCC are therefore intrinsically intertwined. Moreover, CMIP data are used by researchers and policy-makers that assess the global and regional impact of climate change.

Through its phases, CMIP data have grown rapidly in volume, exceeding the capabilities of a single institution to handle the collection and distribution. The CMIP6 data volume is expected to exceed the order of 50 PB over its whole phase. Therefore, a global data infrastructure has been set up by the participating data centers, which is called the Earth System Grid Federation (ESGF[81]). The technical setup of ESGF consists of distributed data nodes, arranged in a tier structure, while organizationally, a governance board has been established to steer the federation development and operation and report to the WCRP.

4.3.2 ENES data life cycle

The IS-ENES infrastructure tries to capture the whole data life cycle from model based data generation to data distribution and processing to data archival and data citation. The core provenance related parts of this life cycle are summarized in the following:

A. Data generation, scientific provenance and data publication: This part characterizes the model based data generation process including A.1) the description of the involved model configurations (scientific provenance based on ES-DOC, https://es-doc.org/), A.2) the homogenization of model data outputs according to agreed upon standards and conventions facilitating their intercomparison as well as A.3) the "publication" of the data on ESGF data nodes and portals. The related use case and requirement are: IS-ENES.U1, IS-ENES.R1. The involved provenance "parties" (agents, entities and activities) are illustrated in the following figure:

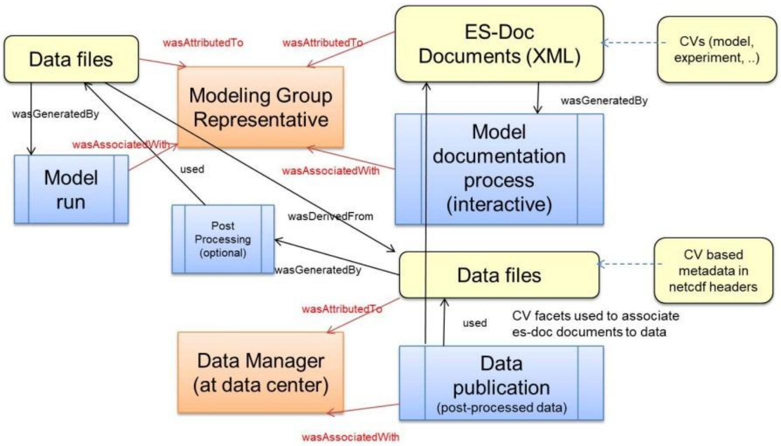

Data files
wasAttributedTo
wasAttributedTo
ES-Doc Documents (XML)
CVs (model, experiment, ..)
Modeling Group Representative
wasGeneratedBy
wasGeneratedBy
wasAssociatedWith
Model run
used
wasAssociatedWith
Model documentation process (interactive)
Post Processing (optional)
wasDerivedFrom
wasGeneratedBy
Data files
CV based metadata in netcdf headers
wasAttributedTo
used
CV facets used to associate es-doc documents to data
Data Manager (at data center)
Data publication (post-processed data)
wasAssociatedWith

Figure 29: Entities related to data generation, model documentation, data publication activities

B) The data generation and ESGF based data publication are often carried out by organisationally and geographically separated entities (modeling centers and data centers) and thus a complex data handover process needs to be performed and documented (data ingest provenance). This

[81] http://esgf.llnl.gov/

handover process generally involves multiples steps: B.1) data submission characterization and announcement B.2) data ingest (including data transmission), B.3) data quality control and finally B.4) ESGF based data publication (corresponds to IS-ENES.U3, IS.ENES.R3). As part of the ESGF data publication persistent identifiers are registered for the data entities and also early citation information is made available.

For the data handover and ingest process no generic, cross data center approach is currently available in IS-ENES, yet data center specific data-ingest provenance information capture and management is supported. The following figure characterizes provenance related actors, entities and activities for the DKRZ data center:

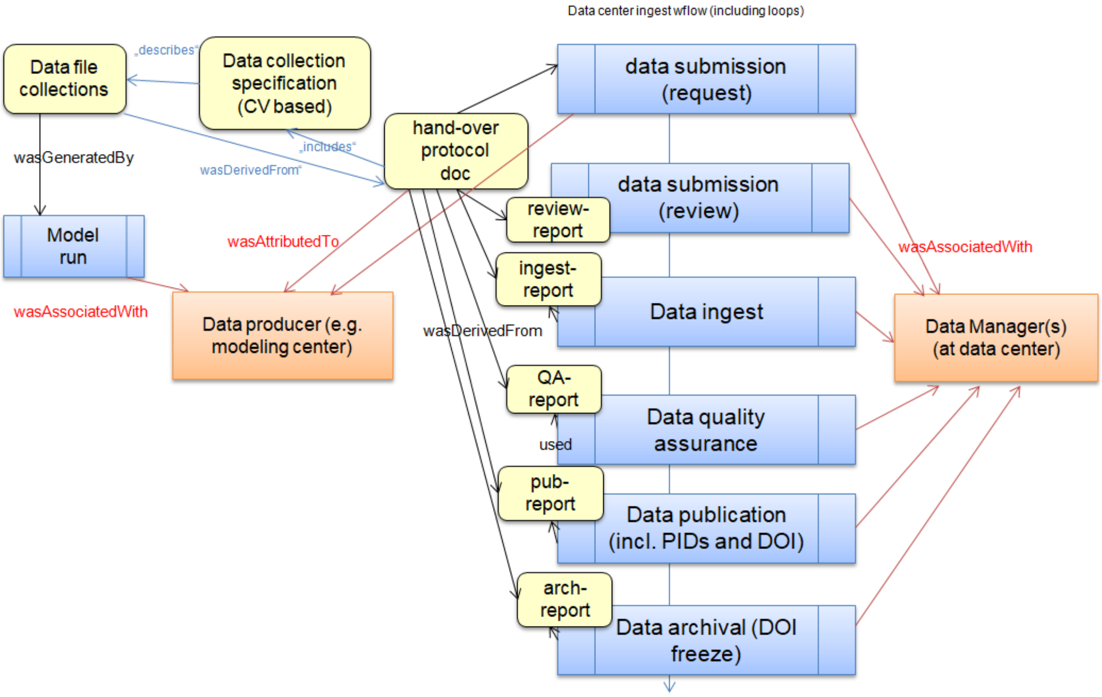


Figure 30: Entities related to data handover to data center

C) After data is published in the European ESGF infrastructure and accessible via the ENES data nodes and data portals errors are detected and thus new data versions and related errata information has to be made available (corresponds to IS-ENES.U4, IS-ENES.R1). The data versioning process as well as the interlinking with errata information is enabled by the underlying (handle system based) persistent identifier infrastructure of the IS-ENES infrastructure: C.1) new versions are published to the ESGF infrastructure and as part of this process new versions are interlinked with old versions in the PID metadata records. C.2) Errata information is collected separately with the help of a github service infrastructure based process and related to the involved data collections based on the PIDs.

D) Frequently aggregate evaluation results based on large amounts of multi-model ensembles of climate data are required and need to make accessible to end users. To use these aggregated results clear provenance records need to be made accessible to document the processing history of these results (IS-ENES.U2, IS-ENES.R2) Infrastructural support for this is currently in discussion and development (see the data typing use case in the next section).

E) As final part of the ENES data life cycle, core data collections are archived and made persistently accessible as part of well curated data centers e.g. the World Data for Climate hosted at DKRZ (IS-ENES.U5). Archival entities are not only the model data collection themselves but also the provenance related entities in the previous data life cycle stages: scientific provenance (es-doc), data ingest provenance as well as data errata and version related information.

4.3.3 Use case: data typing in CMIP6

DKRZ runs one of the tier 1 nodes of ESGF[82], contributes significantly to the ESGF software stack and participates in ESGF governance. Throughout 2016, DKRZ has coordinated and implemented services that will assign a PID (Handle) to every file of CMIP6 and also higher-level aggregates. Also, DKRZ hosts the ICSU World Data Center for Climate, which is tasked with long-term archival of core CMIP6 results, and participates in the IPCC Data Distribution Center.

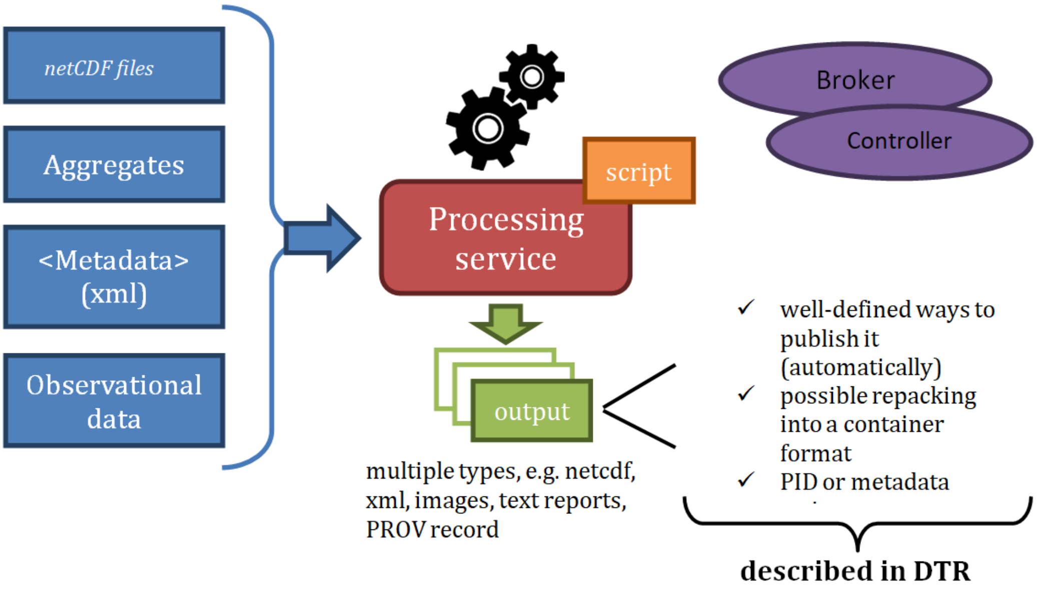

netCDF files
Aggregates
<Metadata> (xml)
Observational data
script
Processing service
Broker
Controller
output
multiple types, e.g. netcdf, xml, images, text reports, PROV record
✓ well-defined ways to publish it (automatically)
✓ possible repacking into a container format
✓ PID or metadata
described in DTR

Figure 31: data typing in CMIP6

DKRZ is also involved in developing the future data services for Copernicus[83], the European Earth observation program. One part of these developments is to provide processing services that are integrated into a common framework that enables end-users to do tasks such as data sub-setting or calculation of climate indicators. Such tasks can require any combination of CMIP6 modelling data and observational data, thereby presenting a challenge for data discovery through brokering, data integration and automated process orchestration through a controller component. DKRZ has already developed processing services which are expected to be expanded and integrated into the Copernicus framework.

The figure above summarizes the proposed scenario for data processing that will benefit from enhanced data typing. A controller is tasked with performing a specific user task and fulfilling processing targets. The controller will task a broker to discover the required data sources and acquire processing services in case of required data type conversions. The processing service is then fed with a script, containing the actual ‘scientific payload’ that will produce meaningful results. Based on the script, the service will use the data sources provided through the broker.

[82] Tier 1 nodes of ESGF currently are: DKRZ, CEDA, IPSL, JPL, LiU, LLNL.
[83] http://www.copernicus.eu/

The broker arranges for data to be retrieved from various sources: netCDF files from the ESGF CMIP6 repository, aggregates of such files, accompanying metadata, observational data from Copernicus sources. The service will produce some form of output, depending on the script, which may include data and metadata products of various formats. The service, the controller or a surrounding framework should then also deal with the output in a way determined by the script and possibly depending on the output data type. For instance, the output may be pushed back to ESGF or other repositories for publication, possibly preceded by a packaging action into a container format. While winding down, the processing service may also assign new PIDs to objects or containers, and possibly attach small metadata items.

Key design considerations for the whole solution are that it should be automated as far as possible and be made transparent to the user to allow provenance to be asserted, even if a detailed provenance report is not produced, based on typed connections between input and output data PIDs and the processing scripts. If the various potential input data entities are typed and also the scripts that can be fed into the processing service, a surrounding execution framework can perform the required orchestration actions automatically. The following workflow further illustrates an exemplary execution scenario, to be jointly executed by controller, broker and processing service:

5. Analyze the scientific script and determine what input variables it requires.
6. Assemble a list of potentially matching data with these variables from CMIP6 and Copernicus data sources.
7. Assert that the data objects in the list contain the variable data in the right format. Discover and execute format converters if necessary.
8. Execute the processing.
9. For each output object, determine through a matching data type description if and how the output should be published, including questions of packaging, PID, provenance and other metadata assignment.

Some of the required elements to develop such a solution are already in place. The PID services implemented in ESGF assign a PID to every file, and include a small set of metadata elements in the corresponding PID record. While ESGF data nodes such as DKRZ are not allowed to modify data after they receive it for publication in ESGF, the necessary pointers to data type records can be easily included in the PID records. In addition, there is a dataset of another climate data activity (Obs4MIPs) that was enhanced with PIDs as part of an earlier demonstrator[84].

Enhancing processing tooling through such a usage of brokering and data typing will have multiple benefits for end-users. There will be less need to manually intervene with the workflows, for instance to convert input formats for processing or output formats for publication systems to accept. As barriers to using data input sources are reduced, scripts may be extended to work with additional sources, and the intrinsic differences of handling modelling vs. observational data may be blurred. On the processing service provider's end, recovery operations could be enabled, such as re-running workflows despite intermediate changes in the data source interfaces or fail-overs to sources matching the same requested data type in case of availability issues. Finally, downstream usage of output data (including further processing) will benefit from the annotations left by the typing effort.

[84] http://hdl.handle.net/10876/ESGF/4ee9d37b-6454-44bf-b3ef-e738b2ecedb4 - note that the actual access to .netcdf files requires OpenID authentication through one of the ESGF sites, e.g. by requesting a user account on http://esgf-data.dkrz.de

4.3.4 IS-ENES provenance support roadmap

Whereas provenance related information is collected and managed along the whole climate data life cycle (e.g. model configuration in data generation process, data versioning and errata in data distribution process as well as long term archival process) the motivation to provide and support standardized (e.g. W3C PROV model based) provenance descriptions comes from the usage scenarios of climate model information in the wider climate community (e.g. climate impact research community and other downstream communities) to support interdisciplinary research. Thus e.g. the IS-ENES climate4impact portal (targeting the climate impact research community) starts to incorporate standards based provenance capture. The same is true for efforts to provide "standard" climate evaluation diagnostics and climate indices. All these efforts are closely related to the "generic processing use case" described in the previous section: data products generated based on large collections of climate model data, which are then used, referenced and cited in downstream research activities need standardized associated provenance records. IS-ENES infrastructure "internal" data provenance information for data products needs to be combined with data processing related provenance to achieve this goal.

5 PROVENANCE RELATED ENVRIPLUS IMPLEMENTATION CASES

This section of the report is about WP 9 use cases which tackle provenance issues from different perspectives [Chen 2017].

The use cases mentioned in this chapter are of two different types:

TEST CASE

Such a test case is built on a new RI service and covers topics of relevance to various WPs, such as instrumentation, data flows and training. Moreover the test case is part of the RI's portfolio of implementation cases.

THE IMPLEMENTATION CASE

ENVRIplus ICT expert work together with the RIs on the RIs description of services they expect from ENVRIplus results. Both ENVRIplus ICT experts and RIs representatives invest in the actual implementation and associated services.

5.1 Provenance-related issues in (dynamic) data identification

This work has been done in the context of the IC1 Data identification and citation[85] lead by Alex Vermeulen and Margareta Hellström from ICOS.

5.1.1 Dynamic data

The term "dynamic data" is typically used to refer to one or more of the following situations [Klump et al., 2015, Rauber, 2016, Hellström, 2017]:

- Errors or omissions in an existing dataset are detected and corrected, necessitating the release of a new version of a data set while deprecating the original one.
- An updated version of data is made available after improvements or modifications in processing and/or quality control. The original version may still be considered usable, or it could be deprecated.

[85] https://envriplus.manageprojects.com/projects/wp9-service-validation-and-deployment-1/notebooks/637/pages/385

- New, previously unavailable data is added to a dataset, without making any changes to existing information. This covers both appending new values to the end of a time series or filling gaps (that might earlier have been marked with "missing value" placeholders.

From this we observe that dynamic data handling and data versioning are closely related topics. The need to be able to unambiguously refer to the different versions suggests that persistent and unique identifiers should be applied to distinguish these different dataset object states. Indeed, good practice recommends that the new version should always receive its own persistent identifier [Klump et al., 2017].

Traditionally, PIDs have been assigned to static files containing snapshots taken before and after a change, sometimes at regular intervals (rather than after each individual update) [Hellström 2017, Klump et al., 2017]. In some cases, the replaced datasets are themselves deleted, although their metadata are typically kept, at least for some time. However, in the interests of scientific reproducibility, an intact copy of the original dataset should be kept, thus making it possible to retrieve it (albeit only after agreement with the repository and/or data producer).

Recent recommendations from the RDA Dynamic Data Working Group [Rauber 2016] instead argue that PIDs should instead be assigned to search queries, which may then be repeated at any given time against the current content of a repository. This approach is based on data catalogues and data stores being able to "remember" any updates or changes made, while keeping older information intact – thus allowing a kind of "time machine" functionality. To support the latter, detailed records of relevant temporal aspects, including 1) the timestamp referring to the storage and/or latest update of the data values themselves (preferably on the level of individual data values!); 2) any time range specified in the search request; and 3) the time at which the search request was made. All of these should, if possible, be noted in the provenance of the data set returned after executing the query. (See, e.g., the description of ENVRIplus Implementation Case IC_01 in [Chen 2017].)

Regardless of the underlying technology and procedures chosen for handling dynamic data, it seems clear that collecting and storing adequate provenance metadata is crucial. Items that should be covered include:

- Timestamp of update and/or version of the digital object
- Version number
- Reason(s) for update/change
- Pointer (PID) to the replaced version (to be added to the new version)
- Pointer (PID) to the new version (to be added to the original version)

If possible, a complete version history should be made available, at least at the landing page of the latest version of the dataset.

5.1.2 Subsets and collections

Scientific datasets vary greatly in complexity and scope, depending for example on the related science domain, the number of parameters of interest and the traditions or expectations of the main target group of end users. As an example, observational datasets from a given RI might contain time series of all evaluated and quality-controlled fluxes from individual ecosystem station collected during a calendar year, aggregated to half-hourly values and enhanced with meteorological and biosphere variables - all accompanied with corresponding quality flag values.

Depending on the research question at hand, an end user might only be interested in one or two of the variables - say, fluxes of latent and sensible heat (plus associated quality flags) and the corresponding wind direction and air temperature values. If such an extraction is performed for one single station, perhaps limited further by applying a specific time range (only spring and summer), we are dealing with a classical subset of a larger data object. The subset's lineage is easily expressed; in the simplest case by including the search string in the materials & methods section of publications together with a proper citation of the original dataset. As an example, Kump et al. [2017] propose a method using handle templates for accurate referring to subsets.

A more complex scenario ensues if the end user wants to study energy fluxes extracted from all forest sites operated by the ENVRI, and then combined into a single dataset that could be used to train and validate a model for the exchange of energy between the land surface and the atmosphere. This new data object is then a collection of subsets of the original data sources, introducing a need to carefully document the lineage of each variable – not only keeping track of e.g. the persistent identifiers of each contributing dataset, but also maintaining the links to the stations at which the measurements were carried out. The former will of course allow to extract the names and affiliations of the data producers, therefore supporting proper allocation of credit (see, e.g., the discussion of ENVRIplus Implementation Case IC_09 in [Chen 2017]), while the latter – provided all corresponding (provenance) metadata are complete and linked, gives access to station descriptions, including land cover maps and other 'ancillary' data that will aid in the characterization of the context and environment in which the data were collected.

5.2 Connecting the particle formation research community to research infrastructure

This work has been done in the context of the IC17[86] 'Connecting the particle formation research community to research infrastructure' lead by Markus Stocker from TIB/PANGAEA.

5.2.1 Short description

The quantities researchers report in scientific literature, say summary statistics such as 7:00 for the mean duration of a studied phenomenon, are generally the result of complex workflows. While not always obvious from reading the reported materials and methods, such values may be derived from numbers generated by an instrument of an observatory; acquired, curated, and published by a research infrastructure; processed using one or more computational models; and interpreted by a postgraduate student supervised by a postdoc who may ultimately derive the reported summary statistics. In using environmental data for system-level science we have thus much provenance as a side product. Unfortunately, such provenance is seldom recorded systematically. Building on a use case in aerosol science, specifically the study of new particle formation events, this implementation exercise represents one approach for how infrastructure can support the specification and execution of complex workflows "as a service" to research communities.

Particle formation is an atmospheric process whereby at specific spatial locations at local or regional scale, aerosol particles form and grow in size over the course of a few hours. Particle formation is studied for its role in climate change and human respiratory health.

[86] https://envriplus.manageprojects.com/projects/wp9-service-validation-and-deployment-1/notebooks/843/pages/936

To study these processes, particle formation needs to be detected for where and when it occurs. Having detected particle formation, the processes are characterized for their qualities, e.g. duration, growth rate and other attributes. The detection and characterization of atmospheric particle formation relies on the measurement of particle size distribution.

In the context of particle formation research, particle size distribution as measured by a Mobility Particle Size Spectrometer (MPSS) is *observational data* – in other words primary, uninterpreted data. For each day and location, observational data are processed and interpreted to detect and characterize particle formation. Observational data processing and interpretation are carried out by one or more human experts (typically postgraduate students). This constitutes an *in silico* (i.e., performed on computer) and *human-in-the-loop* scientific workflow.

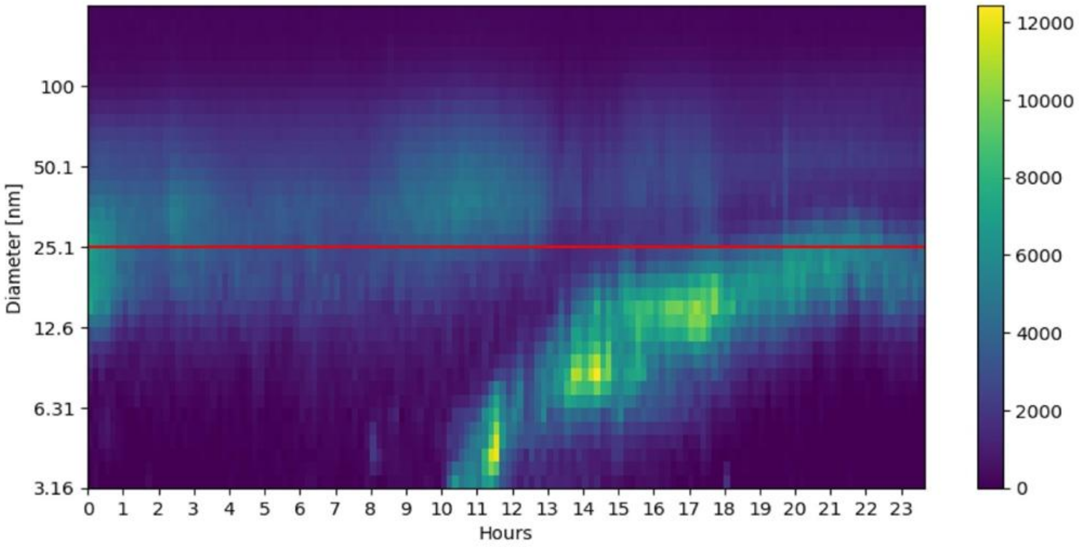
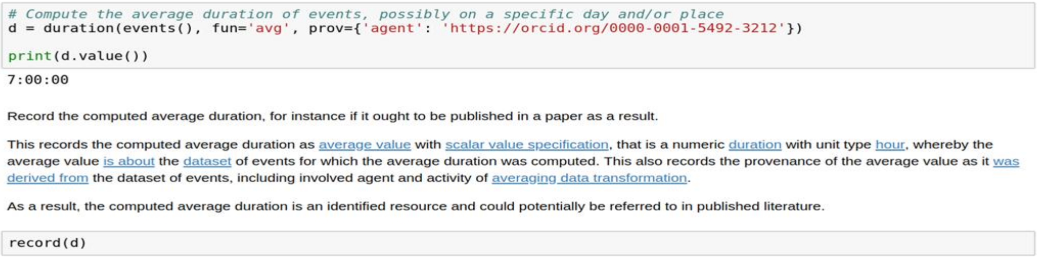


Figure 32: Visualization of primary data in Jupiter notebook

Using a computational environment of their choice, researchers visualize primary data to determine the occurrence of a new particle formation event at the given spatio-temporal locations. The result of primary data interpretation is secondary data describing the event, in particular when and where it occurs, its classification, duration, growth rate and other attributes. Finally, secondary data are used to compute, e.g., summary statistics, such as mean duration of events. Here, mean duration of events is tertiary (computational) data. Such tertiary data may be reported in scientific literature.



```
# Compute the average duration of events, possibly on a specific day and/or place
d = duration(events(), fun='avg', prov={'agent': 'https://orcid.org/0000-0001-5492-3212'})

print(d.value())
```

```
7:00:00
```

Record the computed average duration, for instance if it ought to be published in a paper as a result.

This records the computed average duration as average value with scalar value specification, that is a numeric duration with unit type hour, whereby the average value is about the dataset of events for which the average duration was computed. This also records the provenance of the average value as it was derived from the dataset of events, including involved agent and activity of averaging data transformation.

As a result, the computed average duration is an identified resource and could potentially be referred to in published literature.

```
record(d)
```

Figure 33: Recording of tertiary data

The use case aims to, primarily, (1) harmonize the information describing particle formation; (2) represent information, specifically the meaning of data, using an appropriate computer language; (3) acquire and curate information in infrastructure and (4) develop a concept for provenance and its implementation.

5.2.2 Problems

In the data use phase of the research data lifecycle, researchers currently tend to download data as they are published by research infrastructures onto a local computational environment. We underscore the following problems that come with the practice of download:

- Infrastructural discontinuity: The research infrastructure that publishes primary data is disconnected from the local computational environment. The research infrastructure cannot track what occurs on the local computational environment.
- Systematic recording of provenance: Research infrastructure cannot, in particular, track provenance, which entity is derived from which other entity by what agent and in which activity.
- Heterogeneity of secondary data: As various research groups in the community download and analyse primary data using heterogeneous local computational environments, the resulting secondary data tend to be of heterogeneous syntax and semantics. This is particularly true when the community has not agreed how to represent secondary data using a shared vocabulary.
- Systematic acquisition of secondary and tertiary data: While the curation of primary data, e.g. observational data acquired from environmental sensor networks, is increasingly the responsibility of a professionally managed research infrastructure, secondary and tertiary data are in general not systematically acquired by infrastructure.

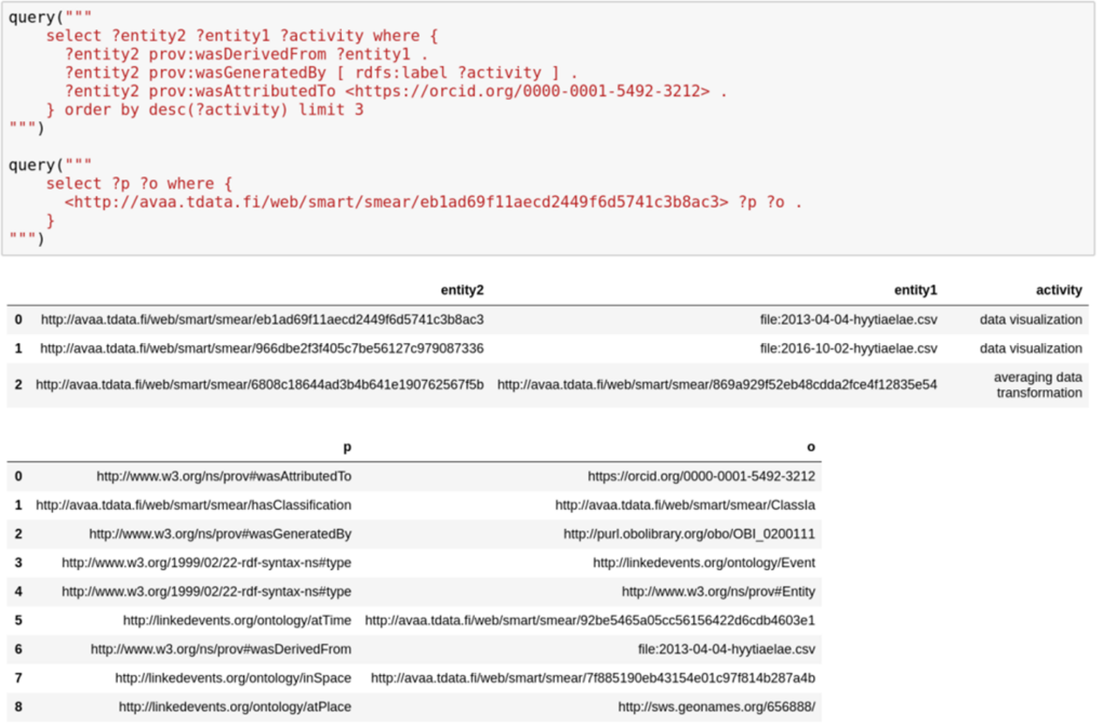
5.2.3 Implementation

For the presented use case in aerosol science, we propose a Jupyter [Pérez and Granger, 2007] based workflow implementation operated “as a service” to the research community on the European Grid Infrastructure (EGI). The implementation addresses the problems underscored above. Operated “as a service,” the federated infrastructure involving both research infrastructure and e-Infrastructure is connected. It avoids (primary) data to be downloaded and is “aware” of the workflows executed. It can thus systematically record provenance. Furthermore, it harmonizes the representation of secondary and tertiary data, specifically descriptions about new particle formation events and computed quantities such as mean duration of events. Finally, secondary and tertiary data are systematically acquired by research infrastructure, guaranteeing the curation and, possibly, the publication of such data, thus enabling their further processing - and the closure of the research data lifecycle. We adopt semantic web technologies and represent secondary and tertiary data in RDF. Following a concept of the Ontology for Biomedical Investigations[87], tertiary data are data items “produced as the output of an averaging data transformation [the activity] and represents the average value of the input data [the entity, here a set of event descriptions]”. Provenance of entities and involved agents and activities is represented using the PROV Ontology [Lebo et al., 2013].

5.2.4 Provenance

One aim of the implementation case is to represent, acquire and curate provenance relating (summary statistics) to information describing particle formation and to particle size distribution observational data, as well as the involved agents (e.g., researchers) and activities (e.g., data interpretation).

[87] http://purl.obolibrary.org/obo/OBI_0000679



```
query("""
    select ?entity2 ?entity1 ?activity where {
      ?entity2 prov:wasDerivedFrom ?entity1 .
      ?entity2 prov:wasGeneratedBy [ rdfs:label ?activity ] .
      ?entity2 prov:wasAttributedTo <https://orcid.org/0000-0001-5492-3212> .
    } order by desc(?activity) limit 3
""")

query("""
    select ?p ?o where {
      <http://avaa.tdata.fi/web/smart/smear/eb1ad69f11aecd2449f6d5741c3b8ac3> ?p ?o .
    }
""")
```

	entity2	entity1	activity
0	http://avaa.tdata.fi/web/smart/smear/eb1ad69f11aecd2449f6d5741c3b8ac3	file:2013-04-04-hyytiaelae.csv	data visualization
1	http://avaa.tdata.fi/web/smart/smear/966dbe2f3f405c7be56127c979087336	file:2016-10-02-hyytiaelae.csv	data visualization
2	http://avaa.tdata.fi/web/smart/smear/6808c18644ad3b4b641e190762567f5b	http://avaa.tdata.fi/web/smart/smear/869a929f52eb48cdda2fce4f12835e54	averaging data transformation

	p	o
0	http://www.w3.org/ns/prov#wasAttributedTo	https://orcid.org/0000-0001-5492-3212
1	http://avaa.tdata.fi/web/smart/smear/hasClassification	http://avaa.tdata.fi/web/smart/smear/ClassIa
2	http://www.w3.org/ns/prov#wasGeneratedBy	http://purl.obolibrary.org/obo/OBI_0200111
3	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://linkedevents.org/ontology/Event
4	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/prov#Entity
5	http://linkedevents.org/ontology/atTime	http://avaa.tdata.fi/web/smart/smear/92be5465a05cc56156422d6cdb4603e1
6	http://www.w3.org/ns/prov#wasDerivedFrom	file:2013-04-04-hyytiaelae.csv
7	http://linkedevents.org/ontology/inSpace	http://avaa.tdata.fi/web/smart/smear/7f885190eb43154e01c97f814b287a4b
8	http://linkedevents.org/ontology/atPlace	http://sws.geonames.org/656888/

Figure 34: Provenance between primary, secondary and tertiary data

Following the PROV Ontology, primary, secondary and tertiary data are entities. In our workflow, mean duration of events (tertiary data) are entities derived from a set of event descriptions (secondary data) which themselves are derived from particle size distribution data (primary data). Various agents and activities are involved, in particular human (researchers) and computational agents and the 'data visualization' and 'averaging data transformation' activities. Relationships between such entities, agents and activities can be acquired, curated and potentially published and processed by infrastructure.

5.2.5 Discussion and Conclusion

We are currently attempting to more actively involve the research community in these developments. Its involvement is essential for a number of reasons. First, the community should agree on how to represent secondary data describing new particle formation events. A first step toward harmonized representation was taken by introducing a relevant concept in the Environment Ontology (http://purl.obolibrary.org/obo/ENVO_01001085). Second, the research community should ultimately adopt the proposed service and perform their data drive science workflows *on* research infrastructure, rather than on local computational environments. These are arguably major steps for this research community, steps that require addressing further issues including the systematic publication of secondary data and the collaborative development and use of software but also the maturity of the approach.

5.3 Quantitative Accounting of Open Data Use

This work has been done in the context of the IC9[88] 'Quantitative Accounting of Open Data Use' lead by Margareta Hellström from ICOS and Markus Fiebig from ACTRIS.

[88] https://envriplus.manageprojects.com/projects/wp9-service-validation-and-deployment-1/notebooks/699/pages/550

5.3.1 Objective

In order to maximize the benefit of public investments into collection of geoscientific data, funding agencies are pushing towards re-use of data for multiple purposes, including re-distribution and commercial use. The re-use of data for other than the original purposes is supposed to facilitate new services, thus generating economic growth. This vision requires increasingly more open data policies. On the other hand, many of these observations are collected by scientifically oriented RIs, where documentation of scientific merit in form of citations or use is paramount for scientists' employment and stations' funding. The objective of this IC is thus to **facilitate quantitatively correct accounting of data use in an open data world.**

To this end, some functionality needs to be in place:

1. Primary identification of **all** data archived in a data centre with suitable, but homogeneous granularity. The granularity needs to be fine enough to resolve authorship of a dataset down to an individual principle investigator (PI).
2. Identified data collections need to include references to the primary identifiers of all data contained in the collection.
3. Services quantifying data use need to refer to the primary identifiers issued by the data's primary data centre as accounting reference. The primary identifiers will usually be DOIs issued by the primary data centre with homogeneous granularity. References to identified data collections need to be resolved to the primary identifiers to facilitate correct quantification of data use resolving the author / principal investigator's contribution.

This use case works on specifying 1) and 2), and works towards initiatives implementing 3).

By functionality as defined under 'Objective":

- The community of atmospheric RIs in ENVRIplus, i.e. ACTRIS, IAGOS, and ICOS, have been identified as reference group for functionality 1. Using its interoperability meetings as forum, the group has agreed to use IC-9's approach for data identification and to implement functionality 1.
- The Research Data Alliance (RDA) has recently established a working group on research data collections[89]. ICOS represents ENVRIplus and IC_9 in this working group, with the aim of including functionality 2 in the specification currently being written by this working group.
- First attempts towards indexing and accounting of data citations have been taken, e.g. by DataCite[90]. Also e-infrastructures such as EGI or commercial indexing services either have voiced their interest or are relevant in this context. Consultations with DataCite by ENVRIplus partners are ongoing, e.g. through the Technical and Human infrastructure for Open Research (THOR) project[91]. This interaction is utilised to investigate options for implementing functionality 3. As a result, DataCite has stated that it will not resolve access events to collection DOIs back to any primary DOIs referenced in the collection DOI. Primary DOIs have varying granularity between scientific domains, making access or citation events of primary DOIs difficult to compare between domains.

5.3.2 Achievements

Achievements by functionality as defined under 'Objective":

[89] https://www.rd-alliance.org/groups/research-data-collections-wg.html
[90] https://www.datacite.org/
[91] https://project-thor.eu/

- Each of the ENVRIplus RI's in the atmospheric domain is currently developing an implementation plan for IC_9 functionality 1, or is already in the process of implementing it.
- The work for lobbying the RDA working group on research data collections for including in its specification a requirement for referencing any primary identifiers of datasets contained in a data collection in the data collection metadata is ongoing.
- While investigating options for implementing functionality 3, difficulties have been encountered. When using the primary identifiers that resolve the contribution of individual PIs as reference for quantifying data use, it is implicitly assumed that the granularity of primary identifiers is homogeneous across scientific domains. However, granularity of primary data identifiers can vary widely across scientific domains. Especially life sciences use finer granularities for primary data identification than most other domains, which will lead to inflation of citation numbers when data collection identifiers are resolved to the primary identifiers. Thus, an alternative approach has been defined. In this approach, the data use accounting service will be provided by the primary data archive, which is supposed to have a homogeneous granularity in its primary DOIs (functionality 1). In order quantify data use, the primary archive will need to query indexing services how often each DOI has been used by itself or as part of a collection. The resulting balance will be comparable at least within the same archive. To provide such a service, indexing services will need to provide a suitable query interface to the primary archive.

5.3.3 Next Steps

For IC_9 functionalities 1 and 2, the implementation work is being continued. Specification and results will be made available to the ENVRIplus community. For functionality 2, the next step would be a reference implementation, which will however be difficult to achieve in the remainder of ENVRIplus

For IC_9 functionality 3, negotiations with indexing services have been started in the context of ENVRIplus WP6. The interfaces needed to access the indexing services' data holdings in order to implement functionality 3 will be part of these negotiations.

5.4 Data acquisition aspects of LTER data life cycle expressed in PROV-O

This work has been done in the context of the IC2 'Provenance' lead by Doron Goldfarb from LTER-Europe.

This section is dedicated to the representation of specific aspects of the LTER Europe data lifecycle using the PROV ontology (PROV-O[92]). The featured aspects cover the life cycle steps between the generation of datasets based on different data collection approaches and their upload to EUDAT B2SHARE[93] via LTER's DEIMS infrastructure, broken down into three major sections. The first section describes the provenance for data collection (observation, measurement, sample based), the second is dedicated to the storage of the collected data into a local database and the respective generation of datasets for export, while the third covers the upload of the exported datasets to B2SHARE.

[92] http://www.w3.org/TR/prov-o/

[93] https://b2share.eudat.eu/

The provenance chain is currently described using the PROV-O elements from the "basic" and the "expanded" term sets. The featured diagrams follow a consistent visual encoding in order to make the identification of the different PROV elements more convenient. The following Figure 35 provides a legend in this regard, featuring the shapes for PROV agents, activities, entities and locations, respectively. The bottom labeling of each shape reflects the expanded PROV class, while the center label represents its respective LTER realization. For the agent and entity classes, the additionally used expanded PROV-O subclasses are listed beneath.

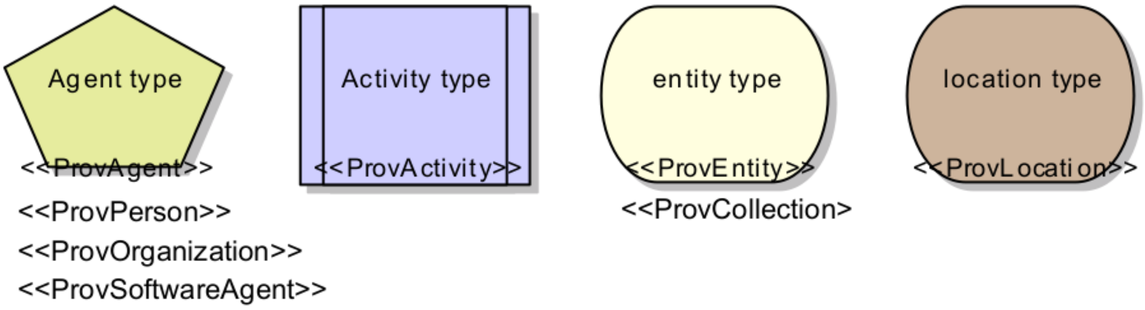


Figure 35: PROV elements used for modeling LTER aspects

Throughout this section, each provenance chain is represented as graph diagram showing the used PROV-O (sub-)classes and their mutual interrelationships using the shapes introduced above. In addition to each diagram, a table lists the individual LTER specific realizations of the different (sub-)classes used in the respective provenance chain.

5.4.1 Data collection

The three main approaches to data collection within the LTER Europe network are based on sensor measurements, sample based methods and direct human observation. For each of these three categories, the recording of provenance obviously differs to quite an extent. This section describes and visualizes these different provenance chains.

5.4.1.1 Sensor based data collection

Sensor based data are delivered from the respective measurement devices. Many of these devices represent complex (digital) machinery with numerous internal processing steps transparent to the user and thus also to the external recording of related provenance. This subsection nevertheless represents an attempt to capture the main data pipeline for such devices in form of PROV in order to assess the vocabulary's expressivity in this regard.

Figure 36 shows an outline for a generic PROV chain for sensor based devices. Starting point is the "sensor" agent on the upper left, representing a specific technical device for detecting some measurable phenomenon. The sensor agent acts on behalf of a "measurement device" agent, which potentially contains multiple sensors. The measurement device itself acts on behalf of a specific "LTER station computer" agent which in turn collects the data from multiple measurement devices. Each station computer eventually acts on behalf of a specific "LTER Site" agent.

A "sensor activity" represents the act of measuring some phenomenon, it takes place at a specific "location of measurement" at a specific time interval. The result of the measurement activity is a "Raw measurement" entity, usually in form of some digitized value. The measurement device stores the individual sensor measurements via a "Recording/Storage" activity, resulting in a “Sensor raw dataset” having the individual measurements as members. The raw dataset is then subject to an automated quality control, e.g. checking for outliers etc. and the controlled dataset subsequently made available to aggregation and processing services such as the calculation of half-hour mean values, performed by an “Aggregation/Processing” software agent. The aggregated and/or processed dataset is represented by a “Sensor dataset with aggregated/processed data” collection at the bottom right, considered to be the final result of the sensor based data collection pipeline.

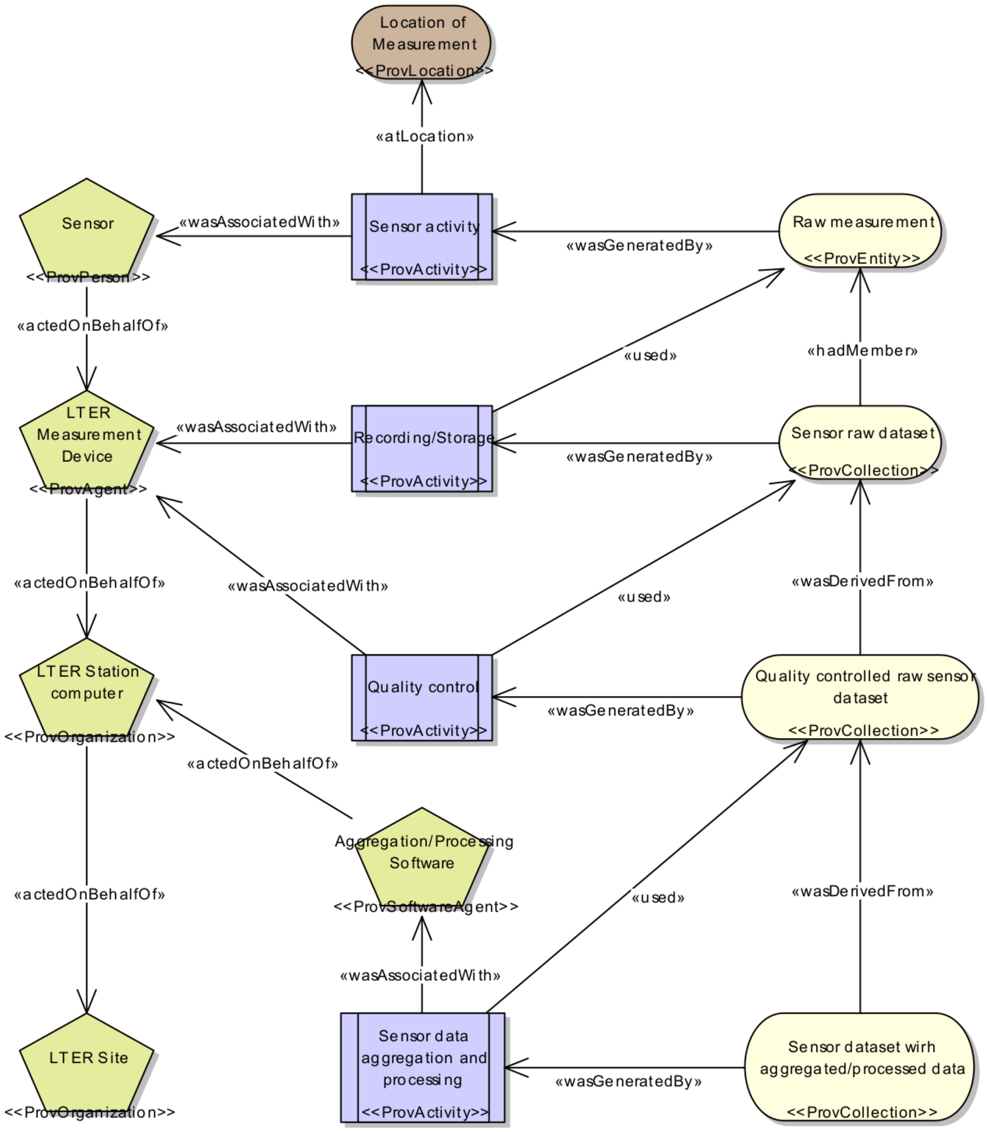


Figure 36: Modeling sensor measurements in LTER via PROV

5.4.1.2 Observation based data collection

In contrast to sensor based measurements, observation based data collection is usually performed by human beings. The basic pipeline for this type of data collection, however, is not significantly different from the machine based counterpart and is shown in Figure 37. Human

"LTER scientist" agents observe a phenomenon at a specific location and point in time. This "Observation activity" results in a respective "Observation entity", e.g. the identification of a specific item within a dedicated area. It is subsequently recorded, for example manually in form of tally marks, resulting in a "Collection of Recorded observations". This artifact is then digitized, i.e. entered into a spreadsheet or database, by a dedicated "Digitization" agent, which can be either human or a machine, resulting in a "Collection of digitized observations". Similar to the sensor based pipeline, this collection is then quality controlled by a dedicated "Quality Assurance (QA)" agent and represents the outcome of this pipeline.

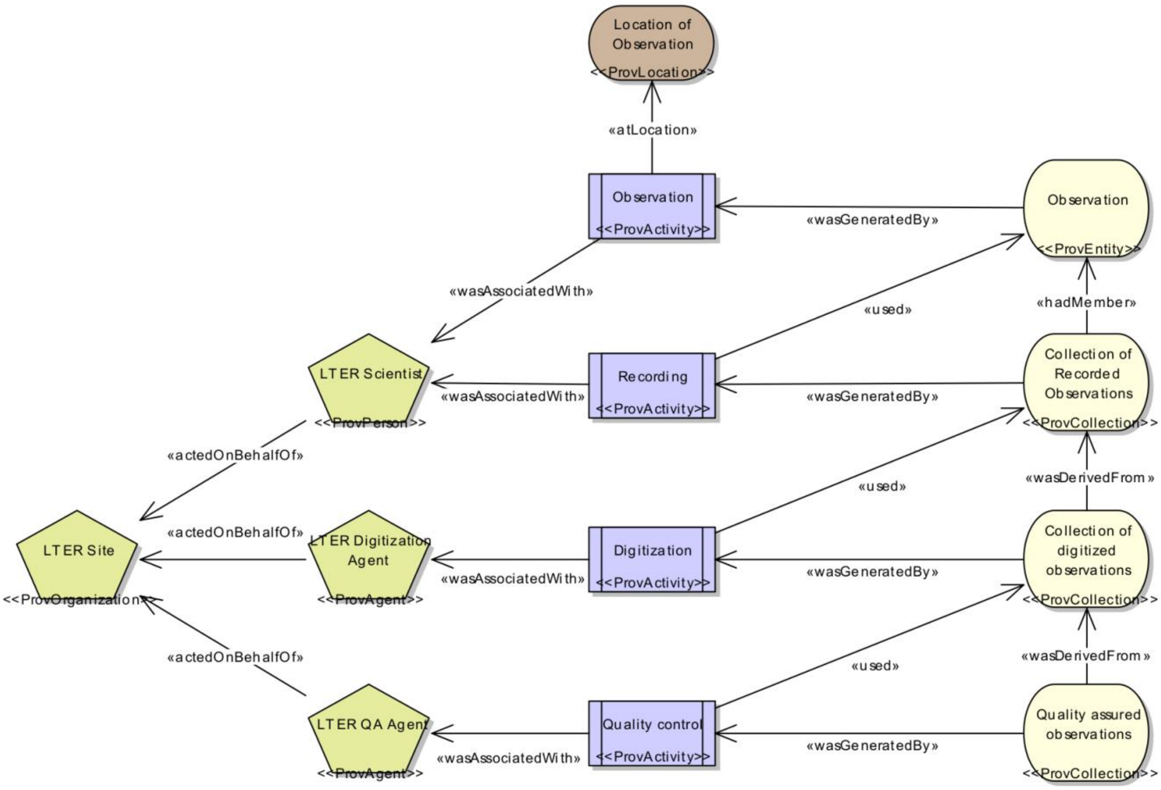


Figure 37: Modeling human observation in LTER via PROV

5.4.1.3 Sample based data collection

Sample based data collection differs significantly from the two previous examples, since it involves physical samples and their lab analysis. Often being mixed from fragments taken at different spots or at different points in time in order to acquire averaged results, samples often have to be treated in very different ways: Some of the substances to be measured are of volatile nature and the respective samples thus need to be processed immediately, while others require expensive lab machinery which is only available in larger facilities. The taken samples are thus often split and processed in different locations and the different results merged together at a later stage.

Figure 38 shows an example for a provenance chain for a sample based data collection. Starting on the upper left, "Scientist" agents perform the "Sampling" activity at a specific location at a specific point in time, resulting in a "Sample" entity. The scientists subsequently mix samples taken at different spots, resulting in a spatially mixed sample. The mixing activity essentially destroys the individual sample entities, which is referred to as “invalidation” in PROV-O. The spatially mixed sample is physically transported to a local laboratory where it is immediately separated into two parts, a “non-volatile” part to be transported to a larger facility and another one for analyzing volatile substances which have to be processed immediately. At this point, the provenance chain thus splits into two separate strands.

In the first strand, the local laboratory processes the sample for volatile substances, yielding a respective result dataset which is made subject to quality assurance. This partial analysis result is currently sent as file but could also be shared with the other parties involved in the analysis via cloud storage services such as EUDAT B2DROP. In the second strand, the local lab first performs a stabilization activity for the “non-volatile” sample part, resulting in a so-called “stabilized sample”. Stabilized samples can subsequently be made subject to a second mixing step which combines samples from different points in time in order to achieve an additional temporal averaging of the already spatially averaged sample. The resulting spatially and temporally mixed sample is subsequently transported to a central lab, where it is first again split into one part dedicated to be preserved and stored for future replications of the analysis, while the other part is analyzed by the lab and the results subsequently made subject to quality assurance. In a final step, the dataset from the volatile substance analysis (e.g. retrieved from B2DROP) is merged with the dataset from the non-volatile analysis, resulting in a combined dataset representing the outcome of the sample data collection process.

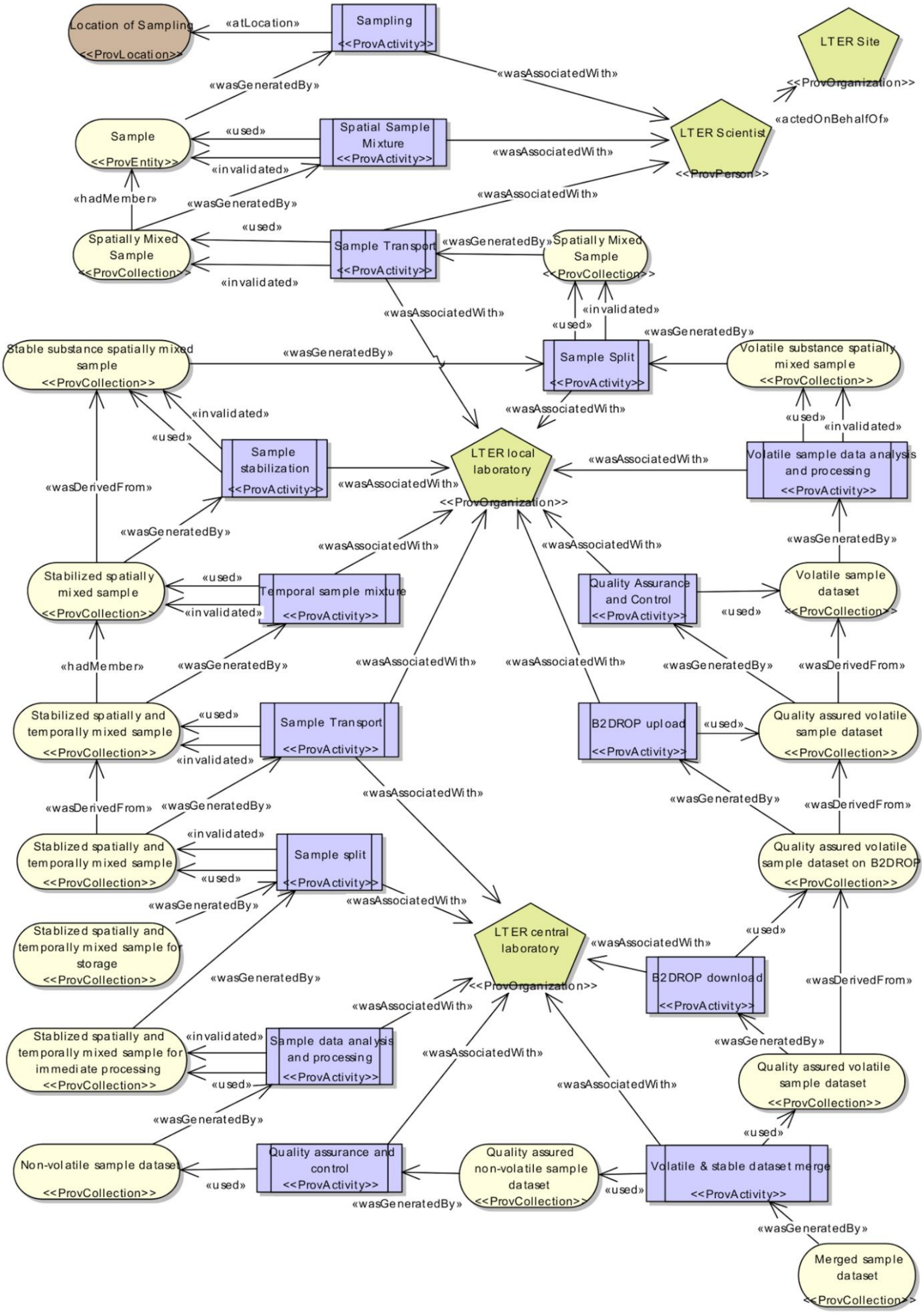


Figure 38: Modeling sample based data collection in LTER via PROV

5.4.2 Data transfer, storage and query/file export

The results from the different data acquisition procedures are usually transferred to a central location and imported into a central database which usually combines the data from multiple sites on institutional or even national level. Such a central DB provides access to different datasets and means for aggregation, querying and exporting. This section considers related provenance chains, separated into the two aspects database import and export, respectively.

5.4.2.1 Import of datasets into central DB

As shown in Figure 39, datasets are transferred to a central computer, where they are imported into a DBMS. Depending on the data collection method, this is done via remote data retrieval or by file transfer. The import of the data includes the data transformation into an appropriate data format. Once imported into the DB, a second quality control takes place in form of a plausibility check, which is performed either manually or automatically by annotating implausible or otherwise suspicious data values.

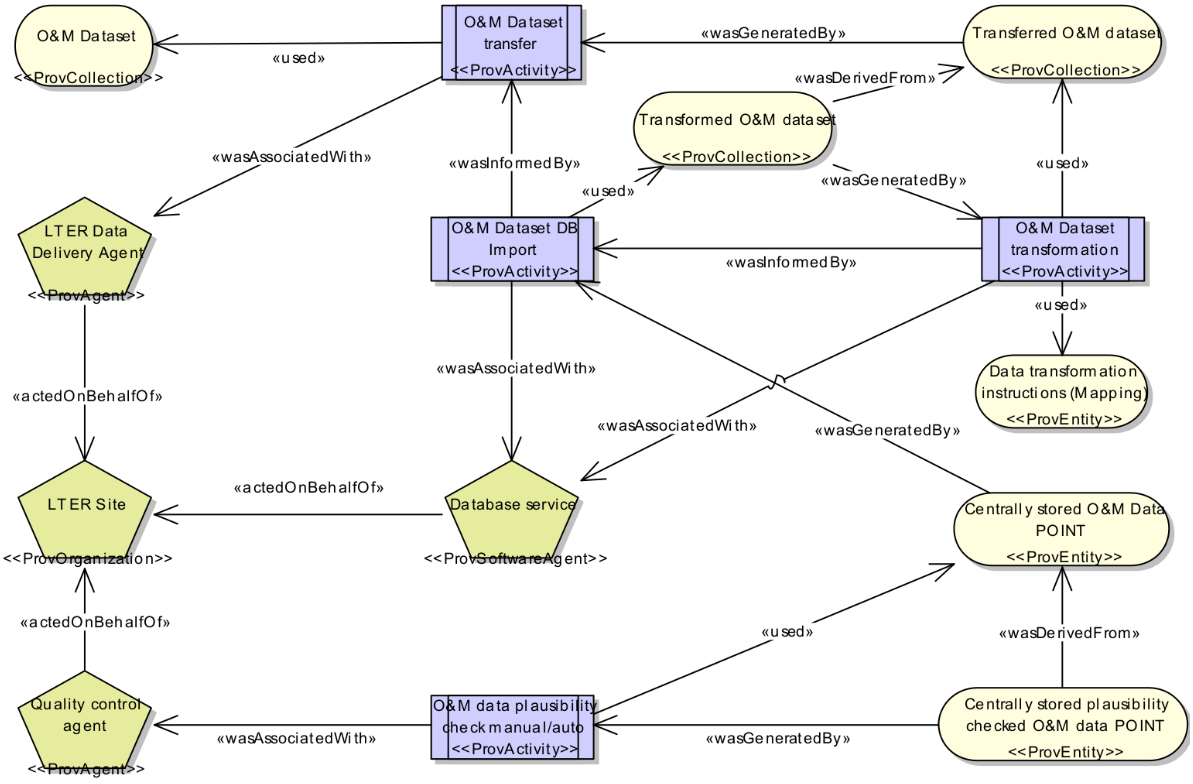


Figure 39: PROV chain for central data storage in LTER

5.4.2.2 Creating datasets for export

The provenance chain for data export, shown in Figure 40, involves the selection, transformation and storage of datasets. An "LTER Data export agent", which can be human or an automated task, performs a selection/query activity on the LTER database which is represented here as PROV-O collection subclass. The query activity is also associated with the respective database service, performing the actual task. Similar to the data import, this again includes the transformation of the extracted data into the required target data format/schema based on explicit transformation instructions. The transformed data is subsequently stored in a file which can be delivered to the intended receivers.

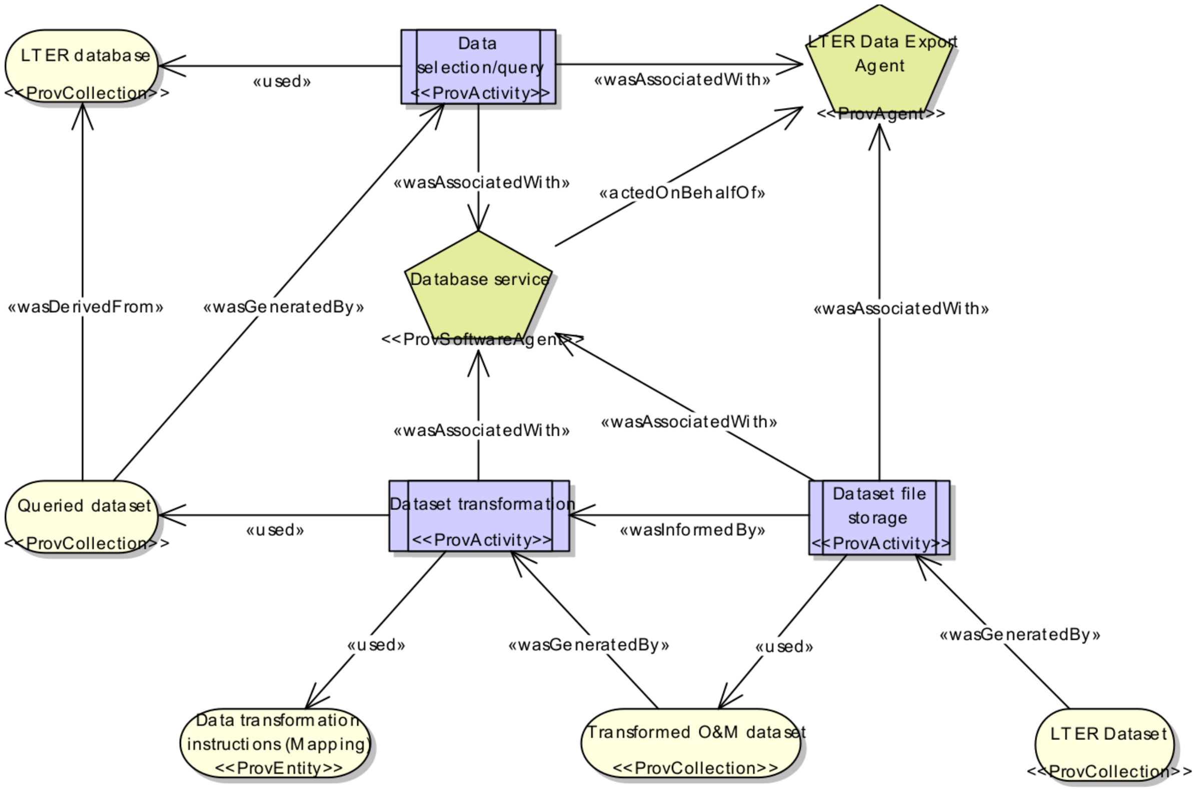


Figure 40: PROV chain for data export in LTER

5.4.3 Data upload/export to EUDAT B2Share

LTER Europe uses the DEIMS-SDR (Dynamic Ecological Information Management System) platform as registry for sites and the respective datasets. In some cases, however, it is also of interest to provide datasets via a larger scale interdisciplinary data hosting infrastructure. Figure 41 provides an overview on an exemplary provenance chain for publishing LTER datasets via DEIMS and EUDAT B2SHARE.

DEIMS requires datasets to be annotated with descriptive information. Datasets and accompanying metadata are subsequently uploaded to and registered in B2SHARE via a dedicated DEIMS (Drupal) module. During registration, B2SHARE generates a unique PID for the dataset using the EUDAT B2HANDLE service. Besides serving as unique ID within B2SHARE, it is also stored in the original DEIMS metadata.

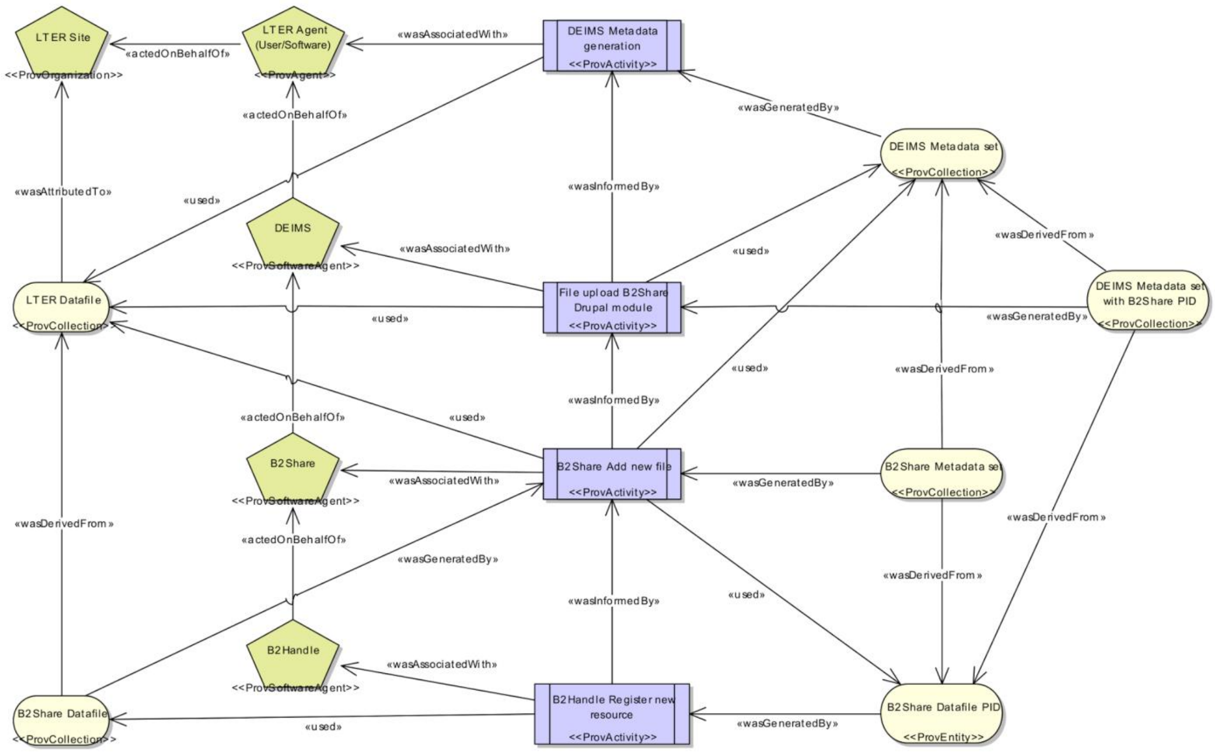


Figure 41: PROV chain for cloud storage upload in LTER

5.4.4 Lessons learned from modeling LTER Data Life Cycle with PROV

Modeling the LTER data life cycle using PROV-O yielded a number of interesting insights. First of all, the modeling process itself required a structural analysis of the dlc aspects, which turned out to be a fruitful exercise in itself, allowing to reflect on the data related processes within the RI in a formal way. PROV-O proved to be a useful framework in this regard, with especially its different levels of expressiveness leaving enough headroom for iterative refinements.

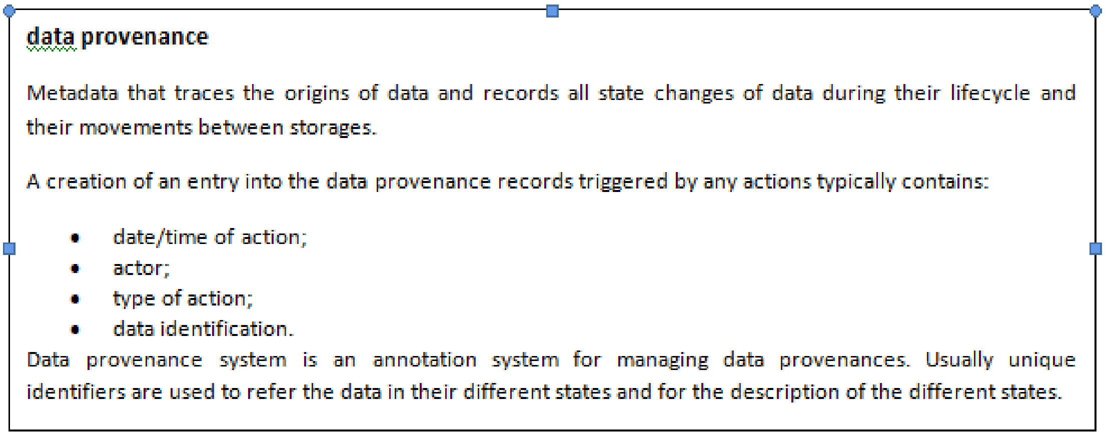
Moreover, the identification of agents, activities and entities during the modeling process immediately raised the requirement for a registry for the respective classes/instances as fundamental building block for provenance tracking systems. At the same time, it made aware of the fact that careful considerations are necessary regarding the required level of detail, otherwise quickly leading to provenance data volumes easily exceeding the actual data.

Next steps will be to refine the resulting PROV models with respect to granularity and expressiveness and to include additional data life cycle aspects such as tracking provenance for script based scientific workflows.

6 RECOMMENDATIONS

6.1 Provenance in the ENVRI RM

The Information Viewpoint is the only section in the current version of the ENVRI RM (V2.2) that refers to provenance in some detail. In it, the information viewpoint defines data provenance as an object type which is a sub-class of the metadata object[94],[95] (Definition 1, Figure 42). This definition of provenance also mentions a provenance system for managing provenance metadata. The Information Viewpoint also defines one action related to provenance management: track provenance[96] (Definition 2).



data provenance

Metadata that traces the origins of data and records all state changes of data during their lifecycle and their movements between storages.

A creation of an entry into the data provenance records triggered by any actions typically contains:

- date/time of action;
- actor;
- type of action;
- data identification.

Data provenance system is an annotation system for managing data provenances. Usually unique identifiers are used to refer the data in their different states and for the description of the different states.

Definition 1: Data Provenance Information Object

The definitions within the Information Viewpoint do not suggest or promoter concrete standards or systems for provenance management, but may include the description of properties, attributes and subtypes, this allows freedom for the design of management, operation and implementation which should be defined at different viewpoint levels (such as the Computational, Engineering or Technology viewpoints).

[94] https://wiki.envri.eu/display/EC/IV+Information+Objects#IVInformationObjects-dataprovenance

[95] https://wiki.envri.eu/display/EC/IV+Information+Objects?preview=/14454557/14454560/IVObjectTypes.png

[96] https://wiki.envri.eu/display/EC/IV+Information+Action+Types#IVInformationActionTypes-trackprovenance

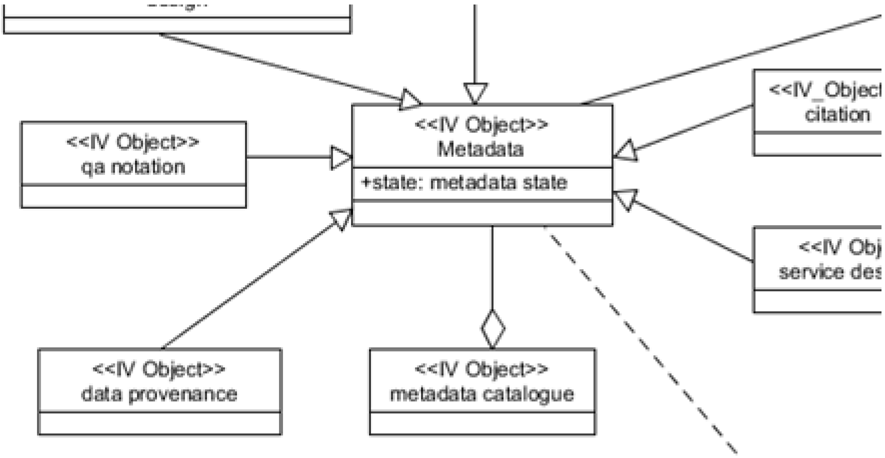
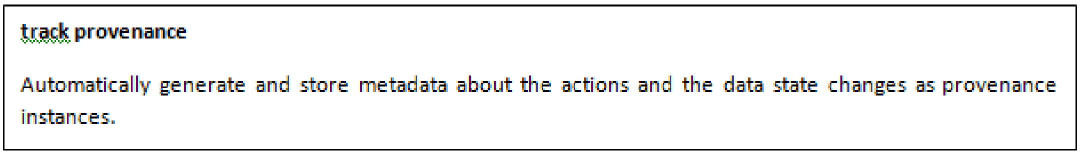


Figure 42: Detail of the Information Object Types Diagram showing the definition of data provenance as a subclass of the metadata object



track provenance

Automatically generate and store metadata about the actions and the data state changes as provenance instances.

Definition 2: Track provenance Information Action

The description of the data lifecycle in the information viewpoint places the “track provenance action” and “data provenance” metadata in the context of the entire data lifecycle as an activity that must be performed whenever an activity that changes the state of a data or metadata object[97] (Figure 43**Fehler! Verweisquelle konnte nicht gefunden werden.**Figure 42). The purpose is to make evident the need to implement provenance tracking as a continuous, parallel activity within the data lifecycle. However, provenance tracking is limited in some areas, for instance use is difficult to track without proper data referencing practices.

The current version of the ENVRI RM (V 2.2) does not further describe specific provenance system components, services, configurations or standards. The main reason for this is that provenance was mentioned briefly and indirectly in one of the original set of requirements for the ENVRI RM[98]. The ENVRIplus project recognised this shortcoming and created a task for the further research into data provenance management and requirements (Task 8.3). The deliverables from that task will serve as reference for extending the coverage of the ENVRI RM in future versions and the corresponding representation in the OIL-E Ontology.

[97] https://wiki.envri.eu/display/EC/IV+Lifecycle+Overview

[98] Definition of the Minimal Model https://wiki.envri.eu/display/EC/Model+Overview#ModelOverview-CommonFunctionswithinaCommonLifecycle and common requirements https://wiki.envri.eu/display/EC/Appendix+A+Common+Requirements+of+Environmental+Research+Infrastructures

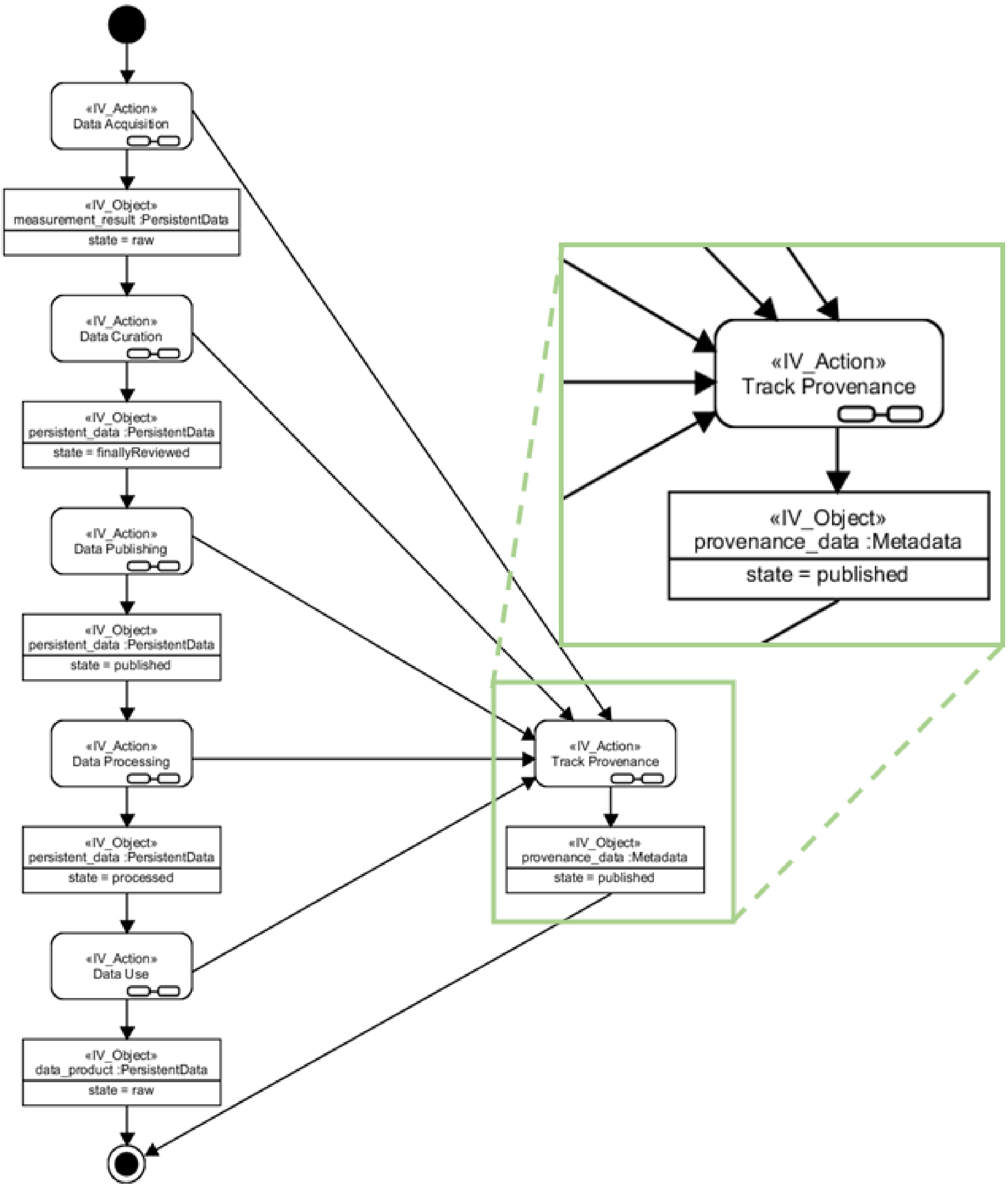

«IV_Action» Data Acquisition
«IV_Object» measurement_result :PersistentData
state = raw
«IV_Action» Data Curation
«IV_Object» persistent_data :PersistentData
state = finallyReviewed
«IV_Action» Data Publishing
«IV_Object» persistent_data :PersistentData
state = published
«IV_Action» Data Processing
«IV_Object» persistent_data :PersistentData
state = processed
«IV_Action» Data Use
«IV_Object» data_product :PersistentData
state = raw
«IV_Action» Track Provenance
«IV_Object» provenance_data :Metadata
state = published

Figure 43: Track Provenance and Provenace Data in the context of the Data Lifecycle

6.2 Provenance and semantic linking

The semantic linking framework of OIL-E provides the ability to create linked data describing research infrastructures, their activities, their component elements and their data. This kind of information has a clear intersection with provenance data, whether as a means to further classify the agents, activities and entities of PROV in terms of ENVRI RM terminology, or simply to describe the provenance services provided by an RI itself. One of the forthcoming actions in this task will be to develop a formal mapping between PROV-O and OIL-E in collaboration with Task 5.3 (semantic linking) in ENVRIplus.

It is expected that the benefit of combining OIL-E and PROV will be twofold:

1. Descriptions of RI activities produced using ENVRI RM and encoded in RDF using OIL-E can be used to verify provenance traces (or alternatively, provenance traces can be used to validate RM-based descriptions of RI activities.
2. An OIL-E knowledge graph can link directly to provenance stores published (or dynamically viewed) in PROV-O.

In the first instance, we can describe the workflow for a behaviour or complex operation within an RI using ENVRI RM concepts, classifying the key actors, artefacts and actions involved at each step in an activity. The RM activity description effectively defines a set of constraints on how that activity is conducted and the actors and entities involved. Given a mapping from PROV to OIL-E (specifically, the ability to classify the actors, entities and activities in PROV to the corresponding concepts in ENVRI RM), those constraints can be applied to a provenance trace to infer as to whether the prescribed process is being correctly followed (for example: are the correct sequence of quality assurance actions being taken and are the right actors involved?). Conversely, this verification process can be flipped to one of validation of an activity description; based on actual traces of activities taken within an RI, it is possible to evaluate as to whether the formal description of a given activity encoded in OIL-E is accurate and correct, by looking for omissions or variations from real traces.

In the second instance, an OIL-E description (which can contain information about specific services or datasets) is directly related to concepts in a PROV-O graph as part of a wider network of linked data. This can simply take the form of RDF triples asserting the classification of certain entities as being of different RM classes in an OIL-E knowledge base, but could also take the form of meta-information about provenance repositories (or even specific provenance traces) themselves.

6.3 Provenance and the ENVRI architecture

Provenance is a critical part of the ENVRI architecture, being key to ensuring the correct attribution of resources to specific RIs (as well as to specific data centres, research groups and individuals working within certain RIs), ensuring the reproducibility of research processes, and engendering trust on the part of researchers in the research outputs of their peers (especially where they might build upon those outputs to produce new research) through the ‘pedigree’ of the asset recorded as provenance.

ENVRI is not an integrated infrastructure, but a loose confederation of research infrastructures that collaborate on shared problems in order to create join solutions. Thus, the purpose of provenance development in ENVRI is not to propose or enforce a single provenance framework within which all research infrastructures should operate, but rather to recommend standard models and approaches to provenance that allow for greater interoperability (through agreed ENVRI - and ideally international - standards for transferring provenance information in a canonical form) between infrastructures. Nevertheless, where shared development takes place, the products of that development should embrace a standard provenance solution that can be interfaced with by all RIs, researchers and other stakeholders in the wider community.

Based on the architecture of the Data for Science theme within the ENVRIplus project, there are a number of pillars of development which contribute to or require a contribution from a provenance framework:

- **Identification and citation.** The persistent identifiers assigned to data collections and other resources provide the preferred way to refer to entities (tools, services, people, sites, etc.) involved in various forms of activity, and thus in provenance traces. It is important that the relationships between the digital objects (assets) are recorded such as: a new version generated by a particular piece of software executed by a particular person.

- **Curation**. Curation activities should include provenance management; provenance traces can be used to locate resources and judge their condition with regard to accessibility and preservation. Conversely, provenance should provide the graph of relationships between curated digital objects (assets).
- **Cataloguing**. The generation of metadata for external (joint) catalogues should be be based partly on provenance records, whether integrated in the source metadata or elsewhere in the source research infrastructure, especially where mapping between metadata standards is involved. The full provenance trace of a given resource should be accessible via any catalogue that contains that resource's metadata. Changes to catalogues should also result in provenance traces that can be used to assess the catalogue themselves. There are particular implications when metadata from a RI catalog is harvested into a common catalog – in itself a provenance action - but also requiring the provenance traces to be harvested using the ENVRI canonical format.
- **Processing.** All activities on the part of a common processing platform should be recorded in the provenance trace of the processes themselves and that of any datasets modified or new data created. This is constrained by catalog information concerning rights, licences and appropriate security and privacy constraints.
- **Optimisation (of e-infrastructure).** The optimisation of e-infrastructure is best facilitated by rich meta-information about the resources being accessed or used, including information about datasets (e.g. geo-location). Historical performance information is also of great use for deciding how optimally cache data or provide infrastructure resources; such information can be retrieved for prior activities if accessible via provenance traces for commonly used data or other resources.

The ENVRI ecosystem is built on three levels of research support environment: virtual research environments, research infrastructures and e-infrastructures. Virtual research environments are potential consumers of provenance and, where integrating processing platforms or workflow systems, contributors. Research infrastructures are potential curators of provenance where directly pertaining to the data collections and other research assets under their stewardship. e-Infrastructures are potential generators of process-oriented provenance that should be retained where possible for use by infrastructure optimisation services.

Provenance data is potentially produced in all types of research environment, but data-oriented provenance needs to be preserved close to the catalogues that provide the metadata on data collections such that there is a direct and stable link from catalogues to provenance stores. In the case of joint catalogues for collecting resource information across RIs (for cross-RI search and discovery) such as the EUDAT B2Find catalogue or the planned ENVRIplus catalogue, where metadata is harvested from multiple source catalogues, it is important to consider a few issues:

- The link to provenance data at the source catalogue must be preserved and synchronized with the joint catalogue.
- The very act of harvesting from one catalogue to another is a provenance event that should be recorded, especially transformation from one metadata scheme to another has been performed.
- If a data collection (or other entity) has been brokered by a joint catalogue and some derivative data product is created, then the role of the joint catalogue is part of the provenance of the result for purposes of accounting and the RI owning the original digital object may desire reflection of the derivative product metadata (and possibly a replicate of the digital object itself) within the IT system of the RI.

Thus it is necessary that there should be a joint framework for catalogue provenance that encompasses both the source catalogues provided by RIs and the joint catalogues developed by communities such as ENVRI. This is best facilitated by the use of a common standard for modelling catalogue-oriented provenance, a standard approach on how provenance is associated with resources catalogued, and a standard API (or small API set) for retrieving provenance data such that the joint and source catalogues can be cross-compatible. In brief, it is essential that provenance is integrated tightly with the catalogs – especially the joint catalog – and with curation.

6.4 Provenance in the longer term

The present ENVRIplus project is aimed at ensuring that the RIs are each well-informed about the work of the others and that they benefit from common approaches and methods, with appropriate common software and appropriate cross-cutting software for interoperation. We have an architectural design in D5.5, recommendations for curation and cataloguing in D8.1 and D8.3 and provenance recommendations (the present deliverable) for ENVRIplus need to fit within that framework. Furthermore, it is intended that all these activities are described by and specified by the ENVRI RM (and in time described in OIL-E).

However, working on provenance (and other areas of the project) stimulates ideas for future architectural components beyond the present ENVRIplus. While there are several different approaches – depending on the priorities of those designing the architecture for the requirements of the ENVRI community – we present one approach with an emphasis on provenance in the expectation that the ideas will be incorporated into an architecture for the ENVRI community beyond the present project.

In the following, the UoDman component would be based on the catalog components of D8.3 and D5.5. The WaaS component similarly would be based on the processing components identified in D5.5. Both are defined somewhat formally by the RM. The P3 component is novel and emphasises a way forward for provenance that is beyond current technologies for provenance (and way beyond current implementations in the RIs).

6.4.1 A fundamental transition

In the longer term federations collaborating to deliver a research infrastructure will reach a scale and complexity that warrants accommodation of internal diversity and dynamic variation [Atkinson *et al.* 2018]. As these progress, the socio-technical dynamics will need to be supported by the architecture. We can see this by examining a core of three groups of services: management of the evolving universes of discourse used by the community, support for all of the actions taken by the community and pervasive persistent provenance, that can be accessed via tools and VREs. The governance of each consortium will need to steer the development to support innovation while sustaining stability, and to recognise diversity while establishing sufficient cross-cutting agreements and standards.

There are significant socio-technical changes emerging that make it essential to reconsider collaboration, the sharing of data and therefore the use of provenance. Examples include, the tension between personal-data legislation and the proper attribution of credit and blame affecting organisations' reputations and individuals' careers; or the funders drive for FAIR publication inhibiting ambitions or commercial engagement. Critical human factors that were previously hidden by technical limitations must now be faced directly; they are already well understood in smaller and more coherent contexts by the CSCW research community [Ackerman *et al.* 2013]. The effects of this transition are brought into sharper focus by two trends:

- The path from collecting initial signal data to their interpretation as evidence about the properties and models of the phenomena being studied is becoming more complex as resolution, sensitivity and scope are extended. This means that many disciplines may be involved and many sophisticated components and steps in the processes may be needed[99].
- The challenges modern research faces are also growing more complex. The filtering of hypothesised models in the science domains as these models compose more mechanisms before the observable signals requires increased ingenuity. Perhaps more significantly, today's research also tries to address global challenges, e.g., how to mitigate the impact of natural hazards or global warming. This inevitable requires sustained collaboration across disciplines, in multi-national, multi-organisational and multi-cultural consortia.

For the first of these trends, provenance becomes more crucial to enable examination and revision of the pathways to validate and improve the quality of the science and the evidence it produces. For the second, as lives, societal well-being and economies are at stake, provenance is essential to establish authority, to protect reputations and to limit the impact of inappropriate actions and malevolent intrusions.

For each of these, provenance has to be *pervasive*, i.e., it must gather information about what has been done by systems and people spanning the complete paths from data coming into the scope of the research enterprise to all its uses as evidence for decisions and as inputs into other actions or enterprises. In the complex consortia needed in the second case, achieving the full span for all of the many forms of data and processes involved is particularly challenging and requires research investment. As we explain below, achieving this requires advances in human and organisational behaviour as well as technical innovation. The technical agenda is mainly about deploying established mechanisms in all contexts and providing good tools to facilitate their use – issues of handling distribution, security and scale may emerge. The human and organisational challenge has to be addressed by delivering incentives; evident and immediate benefits to the practitioners involved [Myers *et al.* 2015]. Gaps in the provenance coverage have two deleterious effects: they introduce vulnerabilities and they reduce productivity.

For each of these, provenance has to be *persistent*, i.e., it should be preserved and available for as long as it is needed. It may be used immediately during an activity to support monitoring, steering and automation. It may be employed decades later when a suspicion of an error is being investigated or when a researcher needs to develop confidence in the former result before

[99] The extremities of this change are well illustrated by contrasting two first observations in astrophysics. Fifty years ago, Jocelyn Bell (as she was then) working as a PhD student shared in the construction of a radio telescope to detect quasars, collected the pen-recorder rolls of paper, examined them, annotated them, and pursued data-handling modifications (increasing paper speed at relevant times) until repeated observations at the correct sidereal time verified her interpretation despite the protests that there was no way of modulating such a powerful energy source 11 times per second. Because of her complete mastery of the path from data to inference, pulsars, now called neutron stars, were discovered [Bell-Burnell 2017]. In 2017 gravitational waves were observed by a global collaboration drawing on a great many domains of expertise but using methods that were defined and monitored by the leading astrophysics and made possible through the intensive sharing of data and processes supported by scientific workflows [Deelman 2017]. Would a PhD student in this context be able to spot and pin down unanticipated phenomena?

building on it. Persistence does not mean ‘just preserving the bits in a store”. It has to mean, being able to interpret the records at a future time with the same understanding and resolution of references as when they were recorded. Some of the referenced data may be too large to preserve for this duration. In such cases a digital tombstone ‘RIP”, with sufficient summary information must replace the original data. Governance will need to decide the criteria and timing of such data terminations. They can be beneficial for the community. For example, sharing the output from large-scale simulations has been shown to be useful in fluid dynamics and astrophysics. This would be unaffordable if that sharing implied long-term support for the model run output, as it may be extremely large. Hence, above a certain size or when overtaken by an improved result, governance may mandate such data termination. If there is any threat of cyberattack or risk of internal misbehaviour the integrity of the persistent record must be protected. Again, governance will rule on the threat level and the choice of counter measures.

6.4.2 Enabling professional judgement and responsibility

A directed imposition of provenance mechanisms would be unacceptable as the professionals[100] would feel inhibited or spied upon and would not cooperate. In many cases, they would migrate to other contexts or hide their work from the system and governance functions – such hidden activity is called ‘skunk work”. The strong feelings are illustrated by comments from professional scientists, e.g.:

- From a LIGO leader, ‘We will not accept automatic mapping in our workflows, we need to control everything ourselves, otherwise false positives will be announced.”
- From an astrophysicist, ‘If you introduce recommenders that even hint at where I am looking, (in SDSS data) I will never get another significant publication.”
- From a senior geoscientist, ‘I want to try out new ideas on my laptop on Saturday afternoons, and tell no one until I know they are worth pursuing.”
- From a rock physicist, ‘I agree negative results must be reported to avoid confirmation bias, but when I have spent months developing a wrong idea, that must not be openly reported – it would ruin my career.”

Many professionals want to feel in control. In fact, we need them to be in *well-informed* control. Few research processes are completely automated. Professionals draw on their expertise and experience to steer processes, select sources, choose targets, tune parameters, create visualisations tailored for recipients and judge whether the accumulated evidence justifies communication or some other action.

Current research cultures often support such control and actions by the professional or team *downloading* the data to a system they control. This mostly is a laptop or local computer, but often it is a virtual work context delivered via a cloud service. They then work on it with tools and workflows they have imported and revised, often using locally developed or improved software. When they have results they upload them, or ship them from their workspace, to a more widely shared context. If that context requires metadata and provenance records, they have to create them or locate and translate local records into the prevailing standard form [Myers *et al.* 2015]. Without built-in provenance support, they have to rely on their memory and their lab notebooks, e.g., if they have modified software or revised control parameters at any stage they need to ensure the new versions are uploaded, documented and identified.

[100] We use the term “professional” to span all of those involved from data providers to users presenting evidence from the data in appropriate forms. This includes many roles and disciplines.

The simple response to this is to state that ‘Explicit use of ‘**download**’ should be deprecated’ because:

- In today’s systems those users then have to work in a less-well supported environment without the toolsets and services that could help them. Myers *et al.* [2015] showed that supporting researchers in the early and exploratory work helped them and led to improved quality of curated results.
- It disenfranchises researchers who cannot set up local arrangements and cannot afford the necessary local support.
- Such transfers lose information, as processes unknown to the consortium’s system occur after *download*, and many potential optimisations are then no longer available. For example, a consortium that has promised to limit its GHG emissions from its computing would have an accounting error.
- Download is taken to imply value, but Trani *et al.* [2016] show this is misleading as downloaded files are inspected and discarded. Trani shows that more information in catalogues avoids this and reduces the load on data services and the costs of data transport. Astronomers find this essential [Szalay & Blakeley 2009].
- As data volumes and data-driven workloads increase, data transfers become unaffordable [Pagé 2018].

Fundamentally, download into *unsupported* contexts undermines processes that improve the quality of methods and data. That does not eliminate *download*, e.g., the selection of material from external sources based on criteria to be delivered to the appropriate context for processing is still essential, as data services cannot host all processing and data needs to be integrated from many sites. There is no opportunity to warehouse all the data needed in one location or to bring it under a single ownership. Instead, download has to be packaged in a structured context, just as today’s programming languages package almost all required forms of goto [Dijkstra 1968].

6.4.3 Incrementally introducing pervasive provenance

The challenge is to introduce attractive systems that fully support provenance and to promote their adoption in every working environment used by the diversity of professionals in all their roles throughout their consortium in such a way that:

1. It is immediately and evidently beneficial.
2. That, as far as possible, it enables them to continue their established practices.
3. It is transparent and controllable what provenance information is collected and which of that information is preserved.

It should equip them with a better understanding of their collaborative work, shared information and evolving methods and thereby improve their capacity to apply well-informed judgements.

Such pervasive provenance infrastructure will take a considerable time and investment to design, build, deploy and refine the mechanisms and tools in every working context. These contexts span the modes of interaction, including the problem-solving tools, the web-based VREs and the programmatic and workflow formalisation and steering of methods. It should include research development and experimental contexts. In many of these contexts there are existing provenance systems prototyped or operational, but they need to be integrated to yield a consistent provenance space. Often the tools or automation that provenance can support have yet to be developed or deployed. These tools may be essential in order that professionals see benefits.

It is a matter for leadership and governance to decide which of these contexts or which aspects in those contexts are a priority for a particular consortium. Different consortia will make different choices. If they build on a common framework and adopt common standards, progress in each consortium will be accelerated. It is a matter for the provenance-system providers to ensure that appropriate and agreed controls are provided, and that provenance-powered tools and methods become available making the benefits significant and self-evident. Leadership and governance need to persuade professionals to adopt the new system or to undertake the responsibilities of traversing the pervasive-provenance boundaries. This may require inducements to early adopters, who then validate and demonstrate the efficacy and usability of the new provenance-collecting work environments. Inevitably, these pioneers will need to uncover residual issues that the providers will need to resolve. The number of these issues and the effort to create and sustain the pervasive persistent provenance system will be substantially reduced by adopting standards, by importing and tailoring existing software, services and databases, and by forming strategic alliances with other users of pervasive persistent provenance. This improvement in rate of provision and quality of the provenance-enabled environments will benefit from the adoption of a shared organisational and technical framework. A proposal for an architecture for such a framework follows.

6.4.4 Architectural interdependencies

The architecture needs to support the work of all roles in the federation from governance to users and external critics. Section 4 of [Atkinson *et al.* 2018] identifies usage patterns for many of the roles.

Architectures that achieve speed of implementation as well as limiting tailoring costs are based on pre-existing *frameworks*. These comprise sets of ready-made components that work well together and can be assembled in quantities that meet a federation's needs while being selected and tailored for each federation using straightforward processes that need specified skills and yield predictable results[101]. Understanding the skilled labour requirements for each stage of construction is critical for planning. Well-made frameworks with such predictability are designed, refined and implemented across a sequence of application domains. An example of such a framework is the framework for VREs[102]. The requirements, activities, priorities, constraints and available technologies all continue to evolve. This requires a co-evolution of the delivered systems.

Consequently, we propose a *flexible federation framework* to facilitate the provision of all aspects of information sharing for all the categories of information a federation needs to share and for all the processes a federation needs to support. Inevitably, this is an open and extensible framework, which provides services and composes many existing components, tools and technologies. The governance and operational teams need to monitor the complexity arising from the way they use the framework and pre-existing systems. Governance may prohibit or curtail options when they no longer provide a sufficiently significant benefit, even though this may mean changing working patterns or established practices. They then have to ensure that the transition costs are met, and that the validity of the change is accepted by those affected.

[101] The Engineering Viewpoint (EV) of the provenance system would identify the components needed and their interrelationships as APIs. The Technical Viewpoint (TV) would then identify one or more available standards and solutions for each of these components. These formalisations have not yet been done and therefore the architecture is described here more informally.

[102] VRE4EIC https://www.vre4eic.eu/

The flexible federation framework (F3) [Atkinson *et al.* 2018] being pioneered y the DARE project[103] has three major subsystems:

- *Universe of Discourse (UoD) management (UoDman)*: which organises the foundations of human and computational communication about the topics that a federation works on.
- *Workflows as a Service (WaaS)*: organises all of the actions required by all of the roles within the federation within the constraints and in the contexts that are specified. A broad view of what constitutes a workflow is taken.
- *Pervasive Persistent Provenance (P3)*: provides trustworthy information flow from the system about all activities and thereby supports a wide range of requirements where humans need to investigate and understand what has happened. It also supports replay, automation and optimisation of repeated tasks.

Treating the F3 trio in concert brings advantages in the implementation of the underlying system and in the coherence of the system's presentation to the many different roles who need to interact with it. Figure 44 shows an overview of the F3 trio with the information flows between the three supporting pillars and to the groups of roles.

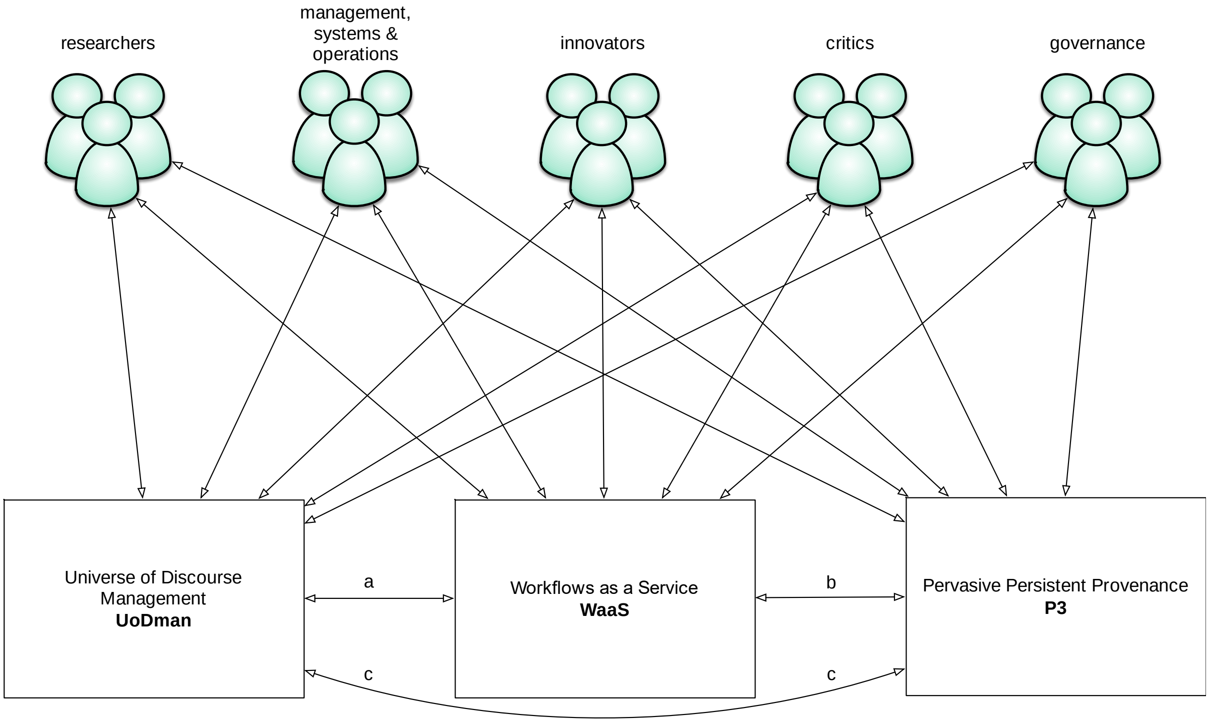

researchers
management, systems & operations
innovators
critics
governance
Universe of Discourse Management
UoDman
a
Workflows as a Service
WaaS
b
Pervasive Persistent Provenance
P3
c
c

Figure 44: The trio of subsystems delivering the Flexible Federation Framework (F3). The three digital pillars provide the foundation for long-term collaboration across evolving multi-faceted federations. All three need to support the diversity and dynamics of the supported community while delivering consistent interpretation wherever and for as long as it is needed.

These different roles within a federation will require different working environments, equipped with the libraries, tools and methods needed by their roles. They will also be subject to different controls, required to use different contexts and be constrained by different rules. There may be significant variations within the groups of roles illustrated above. Some of these require different *digital contexts* encapsulating the trio.

[103] http://project-dare.eu/

We present below more information about the envisaged Pervasive Persistent Provenance (P3) pillar; introducing its scope, function and uses. The other pillars are described in Section 6 of [Atkinson *et al.* 2018] as is the provision of differentiated working contexts. We tabulate P3's uses by each group of professional roles. We identify some of the issues governance should address for this subsystem or group of services as such decisions may raise architectural issues. Different ways of delivering the trio of supporting systems may be chosen in different application domains. Whatever that choice, critical requirements about the relationships between the three pillars of digital support must be met. These are described in terms of the arcs a, b and c in Figure 44.

- The UoDman must provide (a) and preserve all the terms and their definitions the WaaS needs to create and execute workflows. It must provide and preserve (c) for the duration of provenance records all the provenance system (P3) references [Trani *et al.* 2018]. It makes human stated intent more consistent and supports consistent interpretation by maintaining the relationship between human, often context dependent, identifiers, such as names and versions, with the machine interpretable digital identifiers. This has two benefits: supporting translation between those contexts and supporting precise communication by reference. It must also issue provenance records for its own actions. The content required in these communications needs to be clarified, but it should include the relevant metadata and either the relevant data or a sufficiently persistent reference to the data. For example, input data and results may be held in file systems, scientific database systems or reference archives. Whereas, encodings of methods may be held in systems like GitHub, and established terms may be held by ontology services. See section 6.2 of Atkinson *et al.* [2018] for more details.
- The WaaS will request information from the UoDman (a). It may also run workflows to implement UoDman actions (a), e.g., re-building the UoD population to issue a new version, reconciling UoDs that have progressed independently or ingesting new material to add content. The WaaS must supply provenance records for all its actions (b) in the required form, translating if necessary from underlying systems and software. WaaS may conduct actions on the collected provenance data in P3 (b), e.g., to generate summaries or to identify and investigate issues. See section 6.3 of Atkinson *et al.* [2018] for more details.
- The P3 must provide information to support replay, with or without user modification, and restart after partial failures to the WaaS (b). It should be possible to mine the histories of previous runs to support the optimisation of current workflow runs (b) and to enable improved platform management. The UoDman should be able to analyse what changes have been made to a UoDomain (c) in order to generate input to reconciliation and ingest workflows. See section 6.4 of Atkinson *et al.* [2018] for more details.

6.4.5 Professional groups and modes of working

Every federation pooling resources, skills, knowledge and data will need to support all of the groups of roles shown (Figure 45).

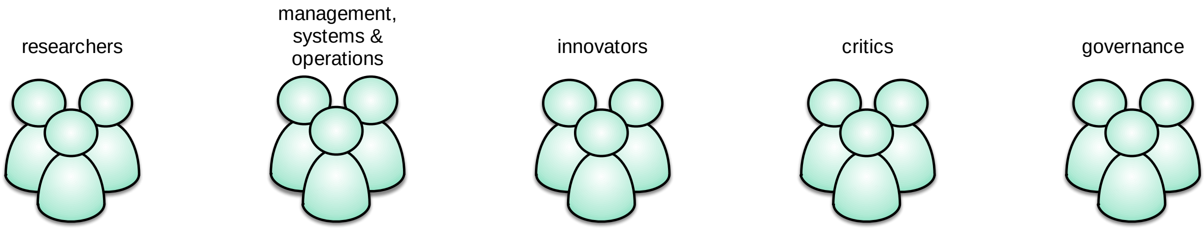


Figure 45: the five clusters of roles that always have to be supported. An individual may work in several roles on different occasions, but they focus on one at any instant. Each group of roles requires tools, methods, and a

work environment that accommodates well the interactions, methods and entities needed for that role. There are two other almost universal clusters of roles that we consider separately: learners and isolated field workers. All of these roles need some facilities and tailoring to match their working practices. But for the federation and its community to function effectively, these tailored working environments need to be consistent and interconnected to promote collaboration between the experts in each of these roles.

Some characteristics of each group of roles are given in Table 16.

Table 16: Characterising dominant clusters of roles in a typical data-sharing community. The work of all of these groups has to be well served by each of the three technology pillars to enable the federation to thrive, be effective and be sustainable in the longer term.

Group of Roles	**Characteristics**
Researchers	To achieve sufficient evidence or to collect data for future work many researchers spend much of their time conducting, refining and quality checking established methods. Students and citizen scientists also mostly work here. Researchers benefit from a stable, well-provisioned work environment that does not disrupt their work by imposing change [Constantin et al., 2018]. They want to make and use their own local refinements. They value familiar tools and adopt changes that improve their own productivity without being disruptive. They worry about the value of their own work and about competition. They have long-running campaigns that require sustained effort and support. They are willing to learn about and use new things only when the perceived benefits outweigh the risk of disruption. 1000s of researchers, 100s of specialisms and 10s of disciplines. In some application domains, these research communities split into almost distinct groups. For example, in climate modelling one sub-group runs the campaigns to build and run better models of the way in which the Earth's climate will change. Another subgroup uses their simulation results as input to explore the ways in which those changes will impact particular local systems, with finer-grained spatial and temporal detail for their chosen focus.
Management, systems and operations	These comprise small numbers of mainly experienced professionals, some with specialist skills, such as systems and software engineering, data architecture and scientific workflow optimisation [Constantin et al., 2018]. Their responsibilities span from immediate response to operational issues and user requests to strategic planning in conjunction with governance decisions. They want to support all roles in their community well by repeating their well-practised operational methods and by deploying their own improvements. They want to do more, with less effort, faster and cheaper if possible. They would like to be more certain about the reliability and value of their work and worry that things they are responsible for may fail. They have to handle emergencies, maintain long-term stability by anticipating requirements and by exploiting new technical and business opportunities.
Innovators	Individuals and teams exploring significant scientific and technological advances that may have extensive pay offs. Maybe 5% of a community or 5% of a professional's time. From research leaders to students in any specialism engaged in the federation. They want to solve major challenges and advance the state of the art by spotting new opportunities and by perfecting new methods. They want as much of the production work environment as they choose with the power to explore changes unfettered by conventions and rules. They need privacy until they are convinced they have an innovation worth sharing. That sharing may depend on further validation, approval and deployment procedures.
Critics	External and internal colleagues, with any role, viewpoint and expertise may investigate work to test its validity or to improve its quality. They may be collaborators, rivals, reviewers and sceptics. This may occur at any time after

	the original work and so requires sufficient provenance information to retain its correct interpretation for a sufficient period. The scale and duration of critical investigations varies greatly. It may consider any aspect of the work of any role, including another critic. This requires that provenance records cover all of the work of all roles in sufficient detail.
Governance	This may be one of the roles of a leader. More commonly, it is undertaken by one or more bodies comprised of trusted representatives of the stakeholders. They may draw on expert advice and action. Their role establishes the moral, ethical and academic ethos, the agreed ways of achieving compliant behaviour and the verification of compliance. This is intended to ensure that the legal and regulatory requirements of each jurisdiction are met, that the treatment of community members is fair, that funders' priorities are addressed, and that resource providers' policies are honoured. When appeals raise issues governance should ensure that they are properly investigated and resolved.
Learners	Learners span the range from novices to experienced experts. They seek to develop or improve their knowledge, skills and judgement for any of the activities in any of the roles. It may include induction into standard practices or introduction to new facilities, capabilities, methods and data. Learners require good simulations of the relevant work environment, with support when requested but privacy so they do not feel they lose face if they do not master challenges rapidly. Their work must persist for them but must not change the state of the public shared environment.
Remote field workers	These predominantly work like Researchers above, though they may turn their hand to any other role when necessary. They usually have a well-defined focus, but once in the field have no, or very limited access, to the normal supporting services.

As we consider each of the three technological pillars, we summarise what each of these groups of roles require from that pillar in a similarly structured table (see Section 6 in [Atkinson *et al.* 2018] for UoDman and WaaS).

6.4.6 Pervasive Persistent Provenance (P3)

The notion of provenance we evoke is a *lingua franca* enabling many different underlying systems, software components and tools to report what they are doing in a form that can be understood throughout the federation and throughout time. It abstracts away from underlying technical detail that shapes logs, error messages and operational telemetry, in order to allow users (individual practitioners, organisations, tools, workflows and services) to interpret consistently the information it contains. As such, it is normally based on internationally adopted standards so that information about actions is consistently represented. These standards started with those required by digital librarians, but the scientific workflow community introduced extensions to capture the dynamics of computational and steered systems. The Kepler workflow group made early progress. After a series of provenance grand challenges, Gill started an activity at W3C with an incubator[104] that proposed a core vocabulary[105] that led to a working group (Moreau and Groth) who saw through the process all the way to W3C PROV standard[106]. That was further developed, driven by the requirements of the DataONE project[107] leading to the

[104] https://www.w3.org/2005/Incubator/prov/charter

[105] https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214

[106] https://www.w3.org/TR/2013/REC-prov-dm-20130430

[107] https://www.dataone.org

ProvONE proposed extension to the standard[108]. This has been further extended, to S-PROV, by Spinuso [2018b], to accommodate and capture aspects of distribution, delegation and runtime changes of workflow execution, with special attention to streaming computational models presenting stateless and stateful operators. Here the concept of *active provenance*, enable runtime oversight and steering [Spinuso 2018b] and the tuneable granularity and precision of the generated provenance traces, thereby allowing for a controllable overhead. Work on the definition, representation and use of provenance continues, e.g., in two RDA working groups[109].

Provenance is supposed to capture whatever information is needed about the actions that have been performed. It should be interpreted consistently in any context where it is read, i.e., it should travel well between organisations and between the different professional roles. Similarly, its interpretation should be consistent throughout time, from the moment it is emitted to support monitoring of active systems to the last time someone wishes to review or re-enact past processes. A tabulation of the provenance provision and other features of 13 popular scientific workflow systems may be found in the appendix of [Atkinson *et al*. 2017].

Different communities may adopt different mappings to these standards, modifying the level of detail, adding domain-motivated additional elements and choosing different durations for preservation. Many application domains have standards that are required for their archived and published data. This may require extractions, summaries and translations from the locally collected provenance records.

Within a community, different work-contexts may require different forms of provenance record. For example, when reviewing progress and consistency in a long-running research campaign visualisations of an abstracted and high-level view of the set of provenance trails may reveal significant patterns or omissions. Researchers may drill down from that view to investigate details drawn from the provenance records. When developing a new method or pattern of working, detailed diagnostic information will help the innovators spot the remaining issues. Often, they want to understand the relationship between data inputs and outputs. Finally, on some occasions, computational performance may have such a high-premium that no provenance records are collected – *not recommended*! The latter can be mitigated by offering users data-driven selective controls to activate provenance recordings when the data meet a specified criterion. This provides professionals with control over provenance-related overheads, for instance, during the execution of real-time streaming workflows. In general, customisable and active contextualisation allows for the rapid analysis of the provenance traces, highlighting the importance of introducing more domain metadata within the traces at runtime [De Oliveira *et al*. 2015]. However, we may argue that such traces would not be sufficiently precise, as they would contain gaps and therefore be misleading for their analysis in the long-term. This aspect could also be addressed by allowing a specific classification of uncertainty of a provenance trace, as anticipated in preliminary work [De Nies *et al.* 2013]. This would still encourage the adoption of provenance-driven solutions, that would allow the incremental refinement of the generated lineage, consistently using the same underlying holistic model and tooling.

[108] https://purl.dataone.org/provone-v1-dev

[109] https://rd-alliance.org/groups/research-data-provenance.html and https://www.rd-alliance.org/groups/provenance-patterns-wg. The latter is developing a database of useful provenance patterns http://patterns.promsns.org

To avoid latency and data-transport costs while distributed workflows are running, provenance may be collected locally. However, logically it is one collection with many references between records. Therefore, it should be treated as and presented as a single information source, recording the history of activity in the infrastructure used by the federation. There are, however, some exceptions to this continuity. For example, for learners looking at their exercise work, the system appears to be integrated with the state presented to them. However, from outside and from other learners' contexts, the provenance associated with their work on the exercise is invisible to preserve learner privacy.

6.4.7 Functions

The functions are relatively straightforward, though the cost of providing software and resources to support them as the federation grows and ages may become significant.

- *Add* additional provenance records.
- *Fetch* specific provenance records given their identities.
- *Find* a set of provenance records that match a query.
- *Filter* a large quantity of entities based on the properties of their ancestors.
- *Traverse* along specified types of arc to additional records.
- *Summarise* a set of records according to a summarisation recipe.
- *Translate* a set of records according to a translation recipe.
- *Ship* a set of records to an external consumer.
- *Reveal* a set of localised records to the shared system.
- *Detect trends in data reuse* often for selected subsets according to a specific set of properties and in the context of the generating methods and workflows executions. This can be applied within focused collaborative campaign as well as in open scientific communities allowing the explorative discovery of intermediate results obtained by peers.

These operations typically handle the transitive closure of referenced records from the identified roots. That may be pruned by *traverse* and by *summarise*.

6.4.8 Uses

The various uses by the groups of roles are summarised in Figure 46. They all use information collected as provenance records to ensure that it has long-term consistency in its interpretation and retains its meaning independent from the diversity and time-varying properties of the underlying systems.

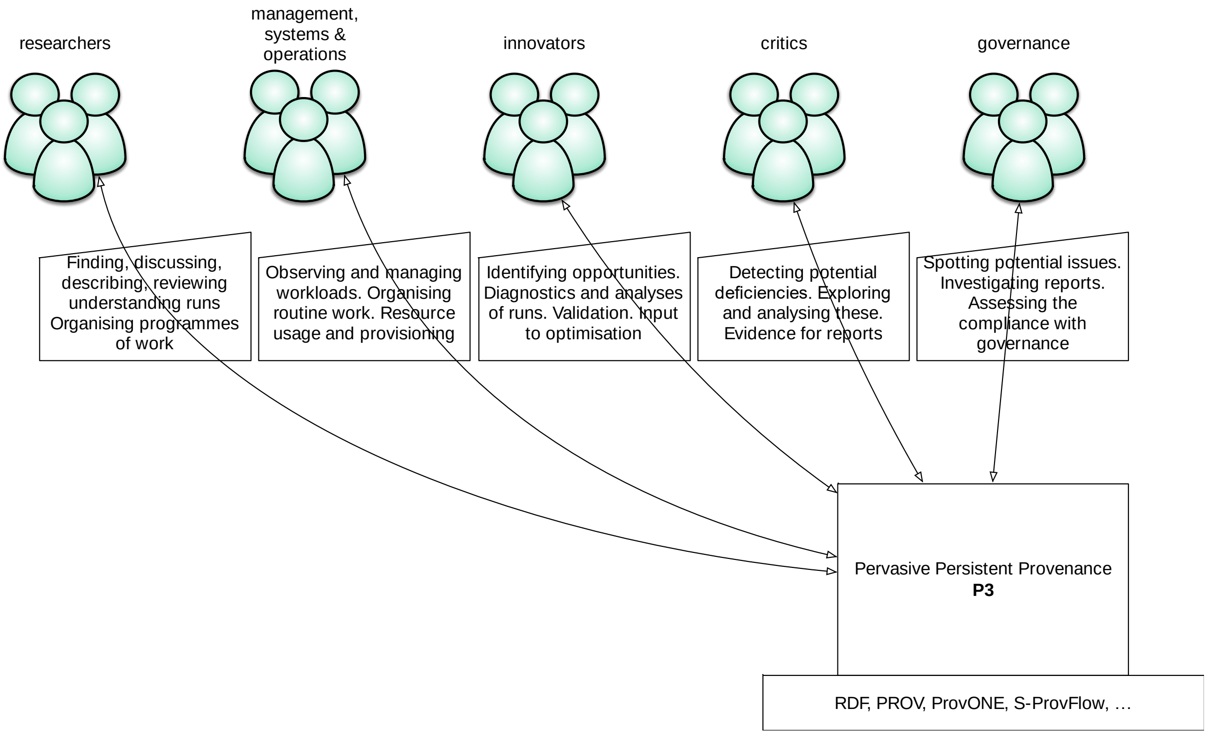

researchers
management, systems & operations
innovators
critics
governance
Finding, discussing, describing, reviewing understanding runs Organising programmes of work
Observing and managing workloads. Organising routine work. Resource usage and provisioning
Identifying opportunities. Diagnostics and analyses of runs. Validation. Input to optimisation
Detecting potential deficiencies. Exploring and analysing these. Evidence for reports
Spotting potential issues. Investigating reports. Assessing the compliance with governance
Pervasive Persistent Provenance
P3
RDF, PROV, ProvONE, S-ProvFlow, …

Figure 46: The uses of Pervasive Persistent Provenance by the principal groups within a federation's community. They use it to better understand their or other's work, to facilitate automation and optimisation, or to observe the overall community and system behaviour. They may also use it to generate metadata that they are required to provide, e.g., for archiving. Such uses depend on good quality tools that interpret the provenance records. The standardisation and stability of provenance representations insulates the provenance users from the details and changes in the underlying systems and software components.

More details of the ways in which each group uses provenance are presented in Table 17.

Table 17: The dominant uses of Pervasive Persistent Provenance (P3) by the principal groups within a federation's community.

Group of Roles	Uses of Provenance
Researchers	These professionals use provenance to explore and examine the processes that have been conducted, e.g., exploring the derivation graph and viewing data inputs that led to selected data items, or the subsequent uses of data items. They find the traces they need to study using queries. They verify that required actions have been performed. They review previous runs and replay them often introducing changes. They manage research campaigns that use many data inputs, run many workflows, involve many people and incrementally build the evidence they are seeking. They conduct searches and analyses, e.g., by running workflows over provenance data, to reveal the points in their campaigns where there are vulnerabilities or recurrent delays. They inject their own domain-relevant metadata into the provenance streams to help them find and navigate provenance information. This domain-oriented metadata is combined with standard provenance information to generate required metadata, e.g., for archiving and curation. Specifying the level of detail in the provenance stream lets them balance provenance collection and handling costs against the potentially available data. However, the use of provenance-summarising tools and visualizers enables them to retain detail without being overwhelmed by it.
Curators and archivists	In many federations some of the researchers in the above group extend their work to include this role. However, some communities have specialists dedicated to these roles. When preparing to curate a data collection passed to them from research or innovation professionals, they verify that the correct

preparation procedures have been conducted by examining the related provenance trails. They may also run workflows to verify quality, promote identifiers and compute fixity signatures. These workflows will use the provenance records to avoid repeating actions already conducted by the providers and to automatically process all of the related objects. For example, a workflow may promote all PIDs to have a global scope and sufficiently durable resolution. The curator's actions will also contribute to the provenance record, so that the quality of curation may itself be examined and improved, and so that corrections may be systematic and as far as possible propagated automatically. As in virtually every role, they will draw on provenance for restart and recovery after partial failures. This should ensure that groups of actions that are logically coupled behave atomically, and are all eventually completed, e.g., registering a PID and preserving both the data and metadata are all completed once one is completed.

Management, systems and operations

Professionals in these roles use the provenance data as a means of observing, monitoring and evaluating the cumulative behaviour and trends in their community's working practices and their system's response to the loads imposed. The lingua franca of provenance allows trends to be studied across subsystem changes as it presents an holistic view of heterogeneous platform components. When operations supports are asked to help with a problem or to investigate a possible incident, the provenance records will be their initial path towards the specifics they need to investigate. They will use restart with modifications to gather more evidence and test their interpretations of what happened. They will monitor the use of data and other resources, and the rate of requests to external providers to decide whether available resources are well used. This will also reveal trends in load that warrant new arrangements, investment in optimisation or planned enhancement to platform elements. The provenance records will be mined for parameters of the prevailing cost functions and to guide optimisation strategies. They will run cumulative usage analyses derived consistently from provenance records to produce reports on usage, productivity and costs in forms required by their community's stakeholders.

Innovators

Innovators pursue new methods, new forms of data and new capabilities for any of the community's roles. They may be specialists in the application discipline or from one of the many areas of supporting expertise. As they develop new methods, they have a dichotomy over their relationship with provenance. On the one hand, they will use it and all of the tools supported by it, that operate in the context they are trying to improve, in order to develop solutions well-tuned to that context and to make use of the information in the provenance trails. They will often increase the detail of provenance collection in order to gather diagnostic and performance detail that would slow production or overwhelm practitioners pursuing routine work. On the other hand, their R&D may involve confronting challenges that take many attempts and much exploration to overcome. They would not want their unproductive paths and repetitions for tests and refinement to be widely viewed. Instead they would like to present externally only the provenance giving the pedigree of their improved products together with evidence that they validated those products thoroughly. Such control may be appropriate for the mainstream pool of shared provenance. However, the behaviour of the innovators should be retained for scrutiny so that security vulnerabilities can be investigated in this context. Otherwise, adopted innovations may provide a route to deep and long-term penetration causing extensive harm.

Critics

Critics will use searches over the provenance records to find the instances that fall within their current focus. They will analyse the associated metadata and

	the graph of related items to refine their understanding of their current issue. They will perform replays (with or without modifications) in order to test their hypotheses about potential weaknesses. They may build workflows that traverse provenance trails in order to investigate trends or to detect discontinuities. Their own work will contribute to the provenance trails and may itself become the subject of criticism or investigation. The pervasive and persistent potential for criticism and investigation is essential for healthy, good-quality science and well-founded evidence-based decision making. Thus, access to the pervasive and persistent provenance records is essential on societal, economic and political grounds. Governance will decide whether it should be granted.
Governance	The governance body has to decide how much to invest in making provenance persistent and pervasive. It needs to decide how carefully provenance needs protecting – cyberattacks or unscrupulous researchers will cover their tracks by overwriting or deleting provenance records if they can. Governance needs to establish the minimum provenance requirements and choose how these are represented, updated and accessed. Governance will analyse the provenance data to assess compliance with the rules it has mandated for the federation, agreed with external providers or must enforce to comply with legislation. It will analyse provenance records to investigate issues and to formulate improved communication mechanisms or rule revisions. It will use the provenance data as an evidential foundation when it handles an appeal.
Learners	As for the other two technology pillars learners may be acting in or preparing for any of a federation’s roles and may be at any stage of maturity and experience. Provenance in the learning context has to be treated specially. People are inhibited if they feel their inevitable mistakes while learning will be revealed to others. Hence, although the presentation to the learner of provenance should be exactly as it would be in the context they are preparing for, including the actions they have taken while learning, that should not be externally visible. This is also required because learners need to know that their actions can do no permanent harm; an erroneous record in a persistent provenance repository could be misconstrued. Most learning will be in a self-managed mode. But a learner may ask for help from a tutor if they reach an impasse. In which case, the learner may grant the tutor access to their lesson’s provenance, so their tutor can review the provenance trail and spot where the learner developed a misunderstanding. The lesson designers may also be granted privileged access to a set of lesson trails (possible pseudonymised) to review and improve the lesson. Finally, if the lesson leads to an element of accreditation or authorisation, the authorising process may verify that all the necessary learning goals were achieved. Otherwise, the provenance generated by each learner as they learn remains local to that learner’s persistent state.
Remote field workers	Field workers will conduct a wide variety of working practices, but each one, for each expedition will be focused on a sub-domain of their community. The P3 system will be locally emulated on their system so that all the provenance-driven tools and methods function correctly. Similarly, the local persistent state will accumulate records of all of the field worker’s activities. When products from the fieldwork are ingested after the return, the corresponding provenance records will also be ingested and merged with the shared P3 records. This may involve some translation, detected during resolution, if the standard representation of the relevant provenance records has evolved during the detached period. The automation of reconciliation and ingest will query the fieldworker’s provenance records to find the set and status of the consistent the new material.

This extensive set of uses will build gradually as the set of provenance-driven methods and tools grows. The effects will vary, from immediate evidence and triggered actions, concurrent with the execution by the system of actions, to long-delayed investigations when an issue is raised, maybe decades later, about the wisdom of certain working practices, the judgement of professionals or the correctness of applied methods.

6.4.9 Summary and implications

Research Infrastructures and collaborative campaigns aspire to last decades during which transformations of methods and approaches will overtake them. This will be driven by technology and provider businesses changing, as well as advances in science and the evolution of the societal and global challenges they address.

The growing wealth of data, with its many sources is opening up many new capabilities for collaborating research consortia and for the communities clustered around one or more environmental research infrastructures. Once the potential for sharing and integrating data, was hidden by the limited digital power available. Today's platforms and software stacks, new algorithms and optimised use of resources have virtually eliminated such digitally imposed limits. The most pressing limits today are socio-economic, their treatment requires innovative leadership developing new strategies for collaborative behaviour. This needs to respect individual and organisational priorities and concerns, be endorsed through governance and be supported by pervasive and persistent computational platforms that have sufficient extent so that they sustain interpretations of methods and data for as long as needed.

A key component of this long-term sustainable research environment is provenance. It acts as the *lingua franca* for communication from diverse and evolving software stacks, tools and services, and from humans – indeed it needs to be capable of recording the *relevant* information about any *action* conducted by any *actor* in a *consistent standardised representation*. It needs to be sufficiently *pervasive* that the record of actions that may be revisited and re-examined is free from informational gaps. It needs to be sufficiently *persistent* that its interpretation remains reliably consistent and unambiguous for as long as needed, maybe decades. The persistent repository needs to support a broad set of functions and uses, through stable interfaces used by tools and VREs. It may also provide a set of tools. If these are stable and consistent across federations, professionals will become skilled in their use and the tools will become well-tuned for each role.

The pervasiveness depends on two interrelated factors:

- The extent to which all of the tools and systems are equipped to collect and transmit appropriate records of action.
- The extent of adoption, which depends on when individuals, groups and projects adopt the provenance-recording working practices and environments.

The later depends on delivering evident benefits, on training, on leadership and on governance.

It is necessary to steer a carefully planned strategy in order to achieve this. That will require investment and commitment as well as good technical support. That sustained investment will be warranted by the improved quality of science built from complex alliances of researchers driving advances in research methods and by the improved scientific productivity enabled by the provenance-powered tools and the provenance-informed investigations and optimisations.

Much of the necessary research has already been accomplished and the methods, technology and standards are ready to use in many contexts and ready to roll out and adopt in others. Development will be necessary to establish consistent implementations for every system, tool and context. Some technical research into how to handle scale and security issues may be needed as this wider adoption occurs. A greater impediment to pervasive persistent provenance comes from human issues. Research is needed to understand how best to address the concerns and worries of organisations and individuals in the context of the large multi-disciplinary, multi-national, multi-cultural consortia needed to sustain campaigns addressing today's complex and urgent social and global challenges. This research needs to focus on the socio-economic and political factors in the context of the rapidly advancing digital ecosystem that is dominated by media, games, business and cyberattacks.

7 REFERENCES

[Ackerman *et al.* 2013] M.S. Ackerman, J. Dachtera, V. Pipek and V. Wulf, 'Sharing knowledge and expertise: The CSCW view of knowledge management', Computer Supported Co-operative Work, 22, 531-573, 2013.

[Altintas et al., 2006] I. Altintas, O. Barney, and E. Jaeger-Frank, 'Provenance collection support in the kepler scientific workflow system', in International Provenance and Annotation Workshop. Springer, 2006, pp. 118–132.

[Atkinson et al., 2015] M. Atkinson, M. Carpen´e, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garth, A. Gemu¨nd, H. Igel et al., 'VERCE delivers a productive e-science environment for seismology research', in e-Science (e- Science), 2015 IEEE 11th International Conference on. IEEE, 2015, pp. 224–236.

[Atkinson et al., 2017] M.P. Atkinson, S. Gesing, J. Montagnat and I. Taylor, 'Scientific Workflows: Past, Present and Future, Future Generation Computer Systems', 75, 216–227, 2017.

[Atkinson et al., 2018] M.P. Atkinson, R. Filgueira, A. Spinuso, L. Trani *et al.* 'Download considered harmful – provokes – a flexible federation framework', DARE technical Report, 2018 URL.

[Asuncion, 2013] H. U. Asuncion, 'Automated data provenance capture in spreadsheets, with case studies', Future Generation Computer Systems, vol. 29, no. 8, pp. 2169–2181, 2013.

[Bailo et al., 2017] D. Bailo, , D. Ulbricht, M. L. Nayembil, L. Trani, A. Spinuso, und K. G. Jeffery, 'Mapping solid earth Data and Research Infrastructures to CERIF'. Procedia Computer Science 106 (2017): 112–121.

[Bechhofer et al., 2010] S. Bechhofer, D. De Roure, M. Gamble, C. Goble, and I. Buchan. 'Research Objects: Towards Exchange and Reuse of Digital Knowledge,' 2010.

[Belhajjame et al., 2015] K. Belhajjame, J. Zhao, D. Garijo, M. Gamble, K. Hettne, R. Palma, E. Mina, O. Corcho, J. M. G´omez-P´erez, S. Bechhofer et al., 'Using a suite of ontologies for preserving workflow-centric research objects,' Web Semantics: Science, Services and Agents on the World Wide Web, vol. 32, pp. 16–42, 2015.

[Bell-Burnell 2017] Dame Jocelyn Bell-Burnell, ‘50 years observing Pulsars’, Royal Society of Edinburgh president’s lecture https://www.rse.org.uk/event/fifty-years-of-pulsars-pulasting-radio-stars/ 2017.

[Bier, 2013] C. Bier, ‘How Usage Control and Provenance Tracking Get Together-a Data Protection Perspective.’ In Security and Privacy Workshops (SPW), 2013 IEEE, 13–17. IEEE, 2013.

[Borkin et al., 2013] M. A. Borkin et al., ‘Evaluation of filesystem provenance visualization tools’, IEEE Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2476–2485, 2013.

[Buneman, 2006] P. Buneman, A. Chapman, J. Cheney, and S. Vansummeren, ‘A Provenance Model for Manually Curated Data’, in Provenance and Annotation of Data, 2006, pp. 162–170.

[Celino, 2013] I. Celino, ‘Human computation VGI provenance: Semantic web-based representation and publishing’, IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 11, pp. 5137–5144, 2013.

[Chen 2017] Y. Chen, B. Grenier, M. Hellström, A. Vermeulen, M. Stocker, R. Huber, B. Magagna, I. Häggström, M. Fiebig, P. Martin, D. Vitale, G. Judeau, T. Carval, T. Loubrieu, A. Nieva, K. Jeffery, L. Candela and J. Heikkinen: ‘Service deployment in computing and internal e-Infrastructures’. ENVRIplus Deliverable 9.1, submitted on August 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf.

[Compton et al., 2014] M. Compton, D. Corsar, and K. Taylor, ‘Sensor Data Provenance: SSNO and PROV-O Together At Last.’ In TC/SSN@ ISWC, 67–82, 2014.

[Constantin et al., 2018] A. Constantin, A. Hardisty, A. Nieva, M. Atkinson, ‘Using Personas as Lenses for a Reference Model’, in Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2018.

[Corke et al., 2016] P. Corke, T. Wark, R. Jurdak, W. Hu, P. Valencia, and D. Moore, ‘Environmental wireless sensor networks’, Proceedings of the IEEE, vol. 98, no. 11, pp. 1903–1917, 2010.

[Costa et al., 2013] F. Costa, V. Silva, D. De Oliveira, K. Ocana, E. Ogasawara, J. Dias, and M. Mattoso, ‘Capturing and querying workflow runtime provenance with prov: a practical approach,’ in Proceedings of the Joint EDBT/ICDT 2013 Workshops. ACM, 2013, pp. 282–289.

[Cox, 2017] S. J. Cox, ‘Ontology for observations and sampling features, with alignments to existing models’, Semantic Web, vol. 8, no. 3, pp. 453–470, 2017.

[De Nies et al. 2013] T. De Nies, S. Coppens, E. Mannens, and R. Van de Walle. ‘Modeling uncertain provenance and provenance of uncertainty’, in W3C PROV. In Proceedings of the 22nd International Conference on World Wide Web, pages 167–168. ACM, 2013.

[De Oliveira et al. 2015] D. De Oliveira, V. Silva, and M. Mattoso, ‘How much domain data should be in provenance databases?’ Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance, page 9, 2015.

[Deelman et al. 2017] E. Deelman et al., ‘Pegasus’ role in gravitational wave detection’, https://pegasus.isi.edu/tag/ligo/ 2017.

[Dijkstra 1968] E. Dijkstra; 'Go To Statement Considered Harmful', Communications of the ACM, 11 (3), 147-148, 1968. doi:10.1145/362929.362947.

[Downs et al., 2015] R.R. Downs, R. Duerr, Goldstein, J.C, Hills, D.J., Parsons, M.A., and Ramapriyan, H.K., 'The Importance of Data Set Provenance for Science,' submitted to Eos Trans. AGU, January 2015.

[Duchein, 1983] M. Duchein, 'Theoretical Principles and Practical Problems of Respect Des Fonds in Archival Science.' Archivaria 16 (1983): 64–82.

[Eyring, 2016] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 'Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization', Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.

[Filgueira et al., 2015] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso, and S. Sanchez-Exposito, 'dispel4py: An agile framework for data-intensive escience', in e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, 2015, pp. 454–464.

[Fleischer & Jannaschk, 2011] D. Fleischer and K. Jannaschk, 'A path to filled archives,' Nature Geoscience, vol. 4, no. 9, p. 575, 2011.

[Gadelha et al., 2012] L. M. Gadelha, M. Wilde, M. Mattoso, and I. Foster, 'MTCProv: a practical provenance query framework for many-task scientific computing', Distributed and Parallel Databases, vol. 30, no. 5-6, pp. 351–370, 2012.

[Garijo and Gil, 2012] D. Garijo and Y. Gil, 'Augmenting prov with plans in p-plan: scientific processes as linked data'. CEUR Workshop Proceedings, 2012.

[Garijo et al., 2014] D. Garijo Y Gil, und O. Corcho, 'Towards workflow ecosystems through semantic and standard representations'. In Proceedings of the 9th Workshop on Workflows in Support of Large-Scale Science, 94–104. IEEE Press, 2014.

[Groth and Moreau, 2013] P. Groth and L. Moreau, 'PROV-overview', W3C, W3C Note, Apr. 2013, http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[Hellström, 2017] M. Hellström, M. Lassi, A. Vermeulen, R. Huber, M. Stocker, F. Toussaint, M. Atkinson and M. Fiebig: 'A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIPLUS strategy to negotiate with external organisations.' ENVRIplus Deliverable D6.1, submitted on January 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf

[Hoekstra and Groth, 2014] R. Hoekstra and P. Groth, 'PROV-O-Viz-understanding the role of activities in provenance', in International Provenance and Annotation Workshop, 2014, pp. 215–220.

[Huynh and Moreau, 2014] T. D. Huynh and L. Moreau, 'ProvStore: a public provenance repository', in International Provenance and Annotation Workshop. Springer, 2014, pp. 275–277.

[Huynh et al., 2016] T. D. Huynh, D. T. Michaelides, and L. Moreau, 'PROV-JSONLD: a JSON and linked data representation for provenance', in International Provenance and Annotation Workshop. Springer, 2016, pp. 173–177.

[Kim et al., 2008] J. Kim, E. Deelman, Y. Gil, G. Mehta, and V. Ratnakar, ‘Provenance trails in the wings/pegasus system’, Concurrency and Computation: Practice and Experience, vol. 20, no. 5, pp. 587–597, 2008.

[Klump et al., 2015] J. Klump, R. Huber and M. Diepenbroek, ‚DOI for geoscience data – how early practices shape present perceptions‘, Earth Sci. Inform., 2015, doi:10.1007/s12145-015-0231-5.

[Kohwalter et al., 2016] T. Kohwalter, T. Oliveira, J. Freire, E. Clua, and L. Murta, ‘Prov viewer: A graph-based visualization tool for interactive exploration of provenance data’, in International Provenance and Annotation Workshop, 2016, pp. 71–82.

[Lebo et al., 2014] T. Lebo, P. West, and D. L. McGuinness, ‘Walking into the Future with PROV Pingback: An Application to OPeNDAP Using Prizms’, in Provenance and Annotation of Data and Processes, 2014, pp. 31–43.
[Liang et al., 2017] L. Jiang, W. Kuhn, and P. Yue, ‘An interoperable approach for Sensor Web provenance’, in Agro-Geoinformatics, 2017 6th International Conference on, 2017, pp. 1–6.

[Lim et al., 2011a] C. Lim, S. Lu, A. Chebotko, and F. Fotouhi, ‘OPQL: A first OPM-level query language for scientific workflow provenance’, in 2011 IEEE International Conference on Services Computing (SCC), 2011, pp.136–143.

[Lim et al., 2011b] C. Lim, S. Lu, A. Chebotko, and F. Fotouhi, ‘Storing, reasoning, and querying OPM-compliant scientific workflow provenance using relational databases’, Future Generation Computer Systems, vol. 27, no. 6, pp. 781–789, 2011.

[Macko and Seltzer, 2011] P. Macko and M. Seltzer, ‘Provenance map orbiter: Interactive exploration of large provenance graphs.’, in TaPP, 2011, pp. 1–6.

[Madougou et al., 2013] S. Madougou, S. Shahand, M. Santcroos, B. Van Schaik, A. Benabdelkader, A. Van Kampen, and S. Olabarriaga, ‘Characterizing workflow-based activity on a production e-infrastructure using provenance data’, Future Generation Computer Systems, vol. 29, no. 8, pp. 1931–1942, 2013.

[Magagna et al., 2018] B. Magagna, M. Atkinson, and M. Stocker, ‘Using data for system-level science: A provenance perspective’, poster at EGU 2018, Vienna.

[McKinley et al., 2017] D. C. McKinley et al., ‘Citizen science can improve conservation science, natural resource management, and environmental protection’, Biological Conservation, vol. 208, pp. 15–28, 2017.

[McPhillips et al., 2015] T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, K. Bocinsky, Y. Cao, F. Chirigati, S. Dey, J. Freire et al., ‘YesWork- flow: a user-oriented, language-independent tool for recovering workflow information from scripts’, arXiv preprint arXiv:1502.02403, 2015.

[Meng et al., 2015] H. Meng, R. Kommineni, Q. Pham, R. Gardner, T. Malik, and D. Thain, ‘An invariant framework for conducting reproducible computational science’, Journal of Computational Science, vol. 9, pp. 137–142, 2015.

[Misra et al., 2008] A. Misra, M. Blount, A. Kementsietsidis, D. Sow, and M. Wang, ‘Advances and challenges for scalable provenance in stream processing systems’, in International Provenance and Annotation Workshop. Springer, 2008, pp. 253–265.

[Missier et al., 2008] P. Missier, N. W. Paton, and K. Belhajjame, ‘Fine-grained and efficient lineage querying of collection-based workflow provenance’, in Proceedings of the 13th International Conference on Extending Database Technology. ACM, 2010, pp. 299–310.

[Missier et al., 2010] P. Missier, B. Lud˙ascher, S. Bowers, S. Dey, A. Sarkar, B. Shrestha, I. Altintas, M. K. Anand, and C. Goble, ‘Linking multiple workflow provenance traces for interoperable collaborative science’, in Workflows in Support of Large-Scale Science (WORKS), 2010 5th Workshop on. IEEE, 2010, pp. 1–8.

[Missier et al., 2013] P. Missier, S. Dey, K. Belhajjame, V. Cuevas-Vicenttin, and B. Ludaescher, ‘D-PROV: extending the PROV provenance model with workflow structure,’ 2013.

[Missier, 2016] P. Missier, ‘The lifecycle of provenance metadata and its associated challenges and opportunities,’ in Building Trust in Information, Springer, 2016, pp. 127–137.

[Moreau et al., 2008] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson, ‘The open provenance model: An overview’, in International Provenance and Annotation Workshop. Springer, 2008, pp. 323–326.

[Moreau et al., 2011] Moreau, Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, et al., ‘The Open Provenance Model Core Specification (v1.1)’. Future Generation Computer Systems 27, Nr. 6 (Juni 2011): 743–56.

[Moreau et al., 2018] L. Moreau, B. V. Batlajery, T. D. Huynh, D. Michaelides, and H. Packer, ‘A templating system to generate provenance’, IEEE Transactions on Software Engineering, vol. 44, no. 2, pp. 103–121, 2018.

[Murta et al., 2015] L. Murta, V. Braganholo, F. Chirigati, D. Koop, and J. Freire, ‘noWorkflow: Capturing and analyzing provenance of scripts’, in Provenance and Annotation of Data and Processes, B. Lud˙ascher and B. Plale, Eds. Cham: Springer International Publishing, 2015, pp. 71–83.

[Nakamoto, 2008] S. Nakamoto, ‘Bitcoin. A Peer-to-Peer Electronic Cash System.’ Online: http://nakamotoinstitute.org/bitcoin/.

[Nieva et al., 2016] Abraham Nieva de la Hidalga, Alex Hardisty and Andrew Jones, ‘SCRAM–CK: applying a collaborative requirements engineering process for designing a web based e-science toolkit’, Requirements Engineering vol. 21, pages 107–129, 2016. DOI 10.1007/s00766-014-0212-0.

[Myers et al., 2015] J. Myers, M. Hedstrom, D. Akmon, and S. Payette. ‘Towards Sustainable Curation and Preservation: The SEAD Project’s Data Services Approach’, 2015.

[Lebo et al., 2014] T. Lebo, P. West, and D. L. McGuinness, ‘Walking into the Future with PROV Pingback: An Application to OPeNDAP Using Prizms’, in Provenance and Annotation of Data and Processes, 2014, pp. 31–43.

[Pagé, 2018] Christian Pagé 'IS-ENES/Climate4Impact Use Case', DARE Kick off Meeting Presentation 2018.
https://drive.google.com/open?id=1HmN1uLgP8egF04d5f4bpoY7aGK8NxnMu

[Pasquier et al., 2017] T. Pasquier, X. Han, M. Goldstein, T. Moyer, D. Eyers, M. Seltzer, and J. Bacon, 'Practical whole-system provenance capture', in Proceedings of the 2017 Symposium on Cloud Computing. ACM, 2017, pp. 405–418.

[Patni et al., 2010] Patni, H., Sahoo, S., Henson, C., Sheth, A., 'Provenance aware linked sensor data.' In: Proceedings of the 2nd Workshop on Trust and Privacy on the Social and Semantic Web. vol. 5, May 2010.

[Pérez and Granger, 2007] F. Pérez, B. E. Granger, 'IPython: A System for Interactive Scientific Computing', Computing in Science and Engineering, 2007, 9(3):21-29. https://doi.org/10.1109/MCSE.2007.53

[Pimental et al., 2015] J. F. N. Pimentel, V. Braganholo, L. Murta, and J. Freire, 'Collecting and analyzing provenance on interactive notebooks: when ipython meets noworkflow', in Workshop on the Theory and Practice of Provenance (TaPP), Edinburgh, Scotland, 2015, pp. 155–167.

[Rauber, 2016] A. Rauber, A. Asmi, D. van Uytvanck and S. Pröll, 'Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use'. Bulletin of IEEE Technical Committee on Digital Libraries, vol. 12, issue 1, May 2016, 6-15. Available at http://students.cs.tamu.edu/ldmm/tcdl/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

[Sahoo, 2009] S. Sahoo, R. Barga, J. Goldstein, A. Sheth, and K. Thirunarayan, ‚Where did you come from... where did you go? An algebra and RDF query engine for provenance,' Tech. rep., Kno. e. sis Center, Wright State University, 2009.

[Sahoo et al., 2011] S. Sahoo, V. Nguyen, O. Bodenreider, P. Parikh, T. Minning, and A. P. Sheth, 'A unified framework for managing provenance information in translational research,' 2011.

[Smith et al., 2005] B. Smith et al., ‚Relations in biomedical ontologies,' Genome biology, vol. 6, no. 5, p. R46, 2005.

[Spinuso et al., 2016] A. Spinuso, R. Fligueira, M. Atkinson, and A. Gemuend, 'Visualisation methods for large provenance collections in data-intensive collaborative platforms', in EGU General Assembly Conference Abstracts, 2016, vol. 18, p. 14793.

[Spinuso, 2018a] A. Spinuso, 'S-ProvFlow and DARE Management for data-intensive platforms', talk at RDA-Europe meeting on data provenance approaches, Barcelona, 15-16th January 2018.

[Spinuso, 2018b] A. Spinuso, 'Active Provenance for Data-Intensive Research', PhD thesis, School of Informatics, University of Edinburgh, submitted 2018.

[Stoffers, 2017] M. Stoffers, 'Trustworthy Provenance Recording using block-chain-like database', Master's Thesis, German Aerospace Center, Köln, 2017.

[Sweeney, 2008] S. Sweeney, 'The Ambiguous Origins of the Archival Principle of" Provenance".' Libraries & the Cultural Record 43, no. 2 (2008): 193–213.

[Szalay & Blakeley 2009] Alexander S. Szalay, José A. Blakeley: 'Gray's laws: database-centric computing in science', The Fourth Paradigm 2009: 5-11.

[Tan, 2017] Y. S. Tan, ‘Reconstructing Data Provenance from Log Files.’ PhD Thesis, The University of Waikato, 2017.

[Trani et al., 2018] Luca Trani, Malcolm Atkinson, Daniele Bailo, Rossana Paciello and Rosa Filgueira, ‘Establishing Core Concepts for Information-Powered Collaborations – pioneered by solid-Earth sciences’, submitted to FGCS.

[Wang et al., 2007] M. Wang, M. Blount, J. Davis, A. Misra, and D. Sow, ‘A time-and-value centric provenance model and architecture for medical event streams’, in Proceedings of the 1st ACM SIGMOBILE international workshop on Systems and networking support for healthcare and assisted living environments. ACM, 2007, pp. 95–100.

[Wang et al., 2016] C. Wang, W. Zheng, and E. Bertino, ‘Provenance for Wireless Sensor Networks: A Survey’, Data Sci. Eng., vol. 1, no. 3, pp. 189–200, Sep. 2016.

[Wilkinson, 2016] M.D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, and P. E. Bourne, ‘The FAIR Guiding Principles for Scientific Data Management and Stewardship.’ Scientific Data 3, 2016.

[Yue et al., 2010] P. Yue, J. Gong, and L. Di, ‘Augmenting geospatial data provenance through metadata tracking in geospatial service chaining’, Computers & Geosciences, vol. 36, no. 3, pp. 270–281, 2010.

[Zhang et al., 2015] Q. Zhang et al., ‘WIP: Provenance Support for Interdisciplinary Research on the North Creek Wetlands’, in e-Science (e-Science), 2015 IEEE 11th International Conference on, 2015, pp. 521–528.

[Zhao et al., 2008] J. Zhao, C. Goble, R. Stevens, and D. Turi, ‘Mining Taverna’s semantic web of provenance’, Concurrency and Computation: Practice and Experience, vol. 20, no. 5, pp. 463–472, 2008.

[Zuiderwijk et al., 2013] A. Zuiderwijk, M. Janssen, and K. Jeffery, ‘Towards an e-infrastructure to support the provision and use of open data’. In Conference for E-Democracy and Open Governement, 259, 2013.

8 APPENDIX A. PROVENANCE REQUIREMENTS TEMPLATE

8.1 Template for use cases and requirements

1 Use case

For each use case, please start a new copy of this form and provide the description in detail and then use the space provided at the end to tell us about any other issues, that you think we may need. Please make use of the tooltip which appears when hovering the pointer over an item.

Use case number ____________________ Name ______________________________

Description ______________________________

Actors involved ______________________________

List of events ______________________________

Entry condition
[optional] ______________________________

Exit condition
[optional] ______________________________

Any other issue
[optional] ______________________________

Supporting materials
[optional] ______________________________

Creator ______________________________

Contributor
[optional] ______________________________

Date ______________________________

Revision
[optional] ______________________________

2 Requirement

For each requirement, please start a new copy of this form and provide the functional requirements in detail and then use the space provided at the end to tell us about any other issues, such as non-functional requirements that you think we may need. Please make use of the tooltip which appears when hovering the pointer over an item.

Requirement number **Name**

Use case number [optional] **Use case name** [optional]

Description

Rationale [optional]

Priority [optional]

Source [optional]

Dependencies [optional]

Conflicts [optional]

Developed prov_practice number [optional] **Name** [optional]

Supporting materials [optional]

Creator

Contributor [optional]

Date

Revision [optional]

Please use this space freely to draw attention to any aspect of the use of the above requirement that may be relevant (not functional-requirements.

Operational

Usability and Humanity

Performance