Project Name	**FREYA**
Project Title	**Connected Open Identifiers for Discovery, Access and Use of Research Resources**
EC Grant Agreement No	**777523**

D4.2 Using the PID Graph: Provenance in Disciplinary Systems

Deliverable type Report

Dissemination level Public

Due date 31 May 2019

Authors Artemis Lavasa (CERN, orcid.org/0000-0001-5633-2459)
Sünje Dallmeier-Tiessen (CERN, orcid.org/0000-0002-6137-2348)
Stephanie van de Sandt (CERN, orcid.org/0000-0002-9576-1974)
Tina Dohna (PANGAEA, orcid.org/0000-0002-5948-0980)
Ketil Koop-Jakobsen (PANGAEA, orcid.org/0000-0002-1540-6594)
Uwe Schindler (PANGAEA, orcid.org/0000-0002-1900-4162)
Christine Ferguson (EMBL-EBI, orcid.org/0000-0002-9317-6819)
Johanna McEntyre (EMBL-EBI, orcid.org/0000-0002-1611-6935)
Frances Madden (British Library, orcid.org/0000-0002-5432-6116)
Simon Lambert (STFC, orcid.org/0000-0001-9570-8121)
Vasily Bunakov (STFC, orcid.org/0000-0003-3467-5690)
Chris Baars (KNAW-DANS, orcid.org/0000-0002-5228-1970)

Abstract This report introduces the work on provenance within the disciplinary pilot applications in FREYA. The various approaches, implementations and future plans or considerations are presented and compared.

Status Submitted to EC 31 May 2019

The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.

FREYA project summary

The FREYA project iteratively extends a robust environment for Persistent Identifiers (PIDs) into a core component of European and global research e-infrastructures. The resulting FREYA services will cover a wide range of resources in the research and innovation landscape and enhance the links between them so that they can be exploited in many disciplines and research processes. This will provide an essential building block of the European Open Science Cloud (EOSC). Moreover, the FREYA project will establish an open, sustainable, and trusted framework for collaborative self-governance of PIDs and services built on them.

The vision of FREYA is built on three key ideas: the **PID Graph**, **PID Forum** and **PID Commons**. The PID Graph connects and integrates PID systems to create an information map of relationships across PIDs that provides a basis for new services. The PID Forum is a stakeholder community, whose members collectively oversee the development and deployment of new PID types; it will be strongly linked to the Research Data Alliance (RDA). The sustainability of the PID infrastructure resulting from FREYA beyond the lifetime of the project itself is the concern of the PID Commons, defining the roles, responsibilities and structures for good self-governance based on consensual decision-making.

The FREYA project builds on the success of the preceding THOR project and involves twelve partner organisations from across the globe, representing PID infrastructure providers and developers, users of PIDs in a wide range of research fields, and publishers.

For more information, visit www.project-freya.eu or email info@project-freya.eu.

Disclaimer

This document represents the views of the authors, and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright Notice

Executive summary

The main focus of this deliverable is the different approaches to provenance as understood, expressed, and implemented by the FREYA disciplinary partners in their various pilot applications. The presentations here outline general approaches to provenance in the particular research context of each organisation, current and future implementations, and describe provenance activities that are supported by persistent identifiers.

This deliverable follows the initial ambition and description of the pilot applications in Deliverable 4.1 and shows the progress that has been made in terms of provenance considerations in these communities. The work described here should be considered the start of a journey. We hope the discourse within and across the pilot applications will be useful for triggering similar discussions within EOSC and in the wider Open Science communities.

Provenance is a key topic in FREYA, both in this Work Package (WP4) and in Work Package 2 (PID Core Services). While the work in WP2 is concerned with provenance of persistent identifier metadata, the work in WP4 focuses on sharing provenance information about resources or the metadata of the resources and their relations with other resources.

The discussion underlined that, despite many solutions on the table, this deliverable is the first building block in our ongoing work on provenance. By examining the current state and future, the discussion about provenance shows that it has the potential to enrich the services and PID Graph with valuable information for the user communities and partnering service providers. Moreover, the discussion helped surface needs and requirements that informed the core service development for the PID Graph.

Contents

1 Introduction

Research provenance, understood as a systematic management of the records of origin of research artefacts is an important aspect of Open Science, as it provides contextual information of how and from what sources research originates. Provenance contributes to FAIR principles (Findable, Accessible, Interoperable, Reusable) as it can facilitate research reusability and reproducibility. Provenance may be able to support findability, too; access to provenance records allows follow-up on the artefacts or actors or events that have been involved in a particular piece of research and use such artefacts for finding other pieces of research having commonalities with the one in question.

With regard to persistent identifiers (PIDs), there are different flavours of provenance that can be discerned: provenance of persistent identifiers themselves and their associated metadata as specific research artefacts, and provenance of other research artefacts with PIDs contributing to making clear statements about the artefacts' origin and connections to other PIDs. Provenance is a core concept for creating connections in the PID Graph to contextualise content persistently. The metadata associated with PIDs helps enrich the PID Graph with the necessary information to support a resource's identity (e.g. contributor/s, production date, etc.), which supports trust.

Outside the context of persistent identifiers, one could distinguish a few different types of provenance. This deliverable mostly focuses on resource provenance and metadata provenance, as they are the ones most prominent in our use cases. The former is concerned with the history of a digital object/artefact/resource and the latter with the history of the metadata itself during the curation process.

FREYA is concerned about all the aforementioned flavours of provenance, with a central service provision in WP2 in the PID providers context (see Deliverable D2.2: "PID metadata provenance", which focuses on provenance for PID metadata), and with WP4 placing emphasis on resource and metadata provenance in disciplinary contexts.

The notion of provenance inevitably varies across research disciplines, and one of the purposes of this deliverable is to capture differences, as well as commonalities, in order to describe available variations in the provenance interpretation and provenance management in a variety of disciplines. Another purpose is to identify the role of PIDs in the research provenance supply, management and reuse.

There are already well-established practices and recommendations concerned with provenance. In terms of standardised means of expressing of provenance, the works of the W3C PROV Family of Documents[1] is worth mentioning. Conceptually, PROV defines the framework of interactions between, and statements about the Agent, Activity and Entity involved. With respect to research, the works of the RDA Research Data Provenance Interest Group[2] and the RDA Provenance Patterns Working Group[3] should be noted.

The stance of FREYA has been to take this world-wide effort into account but not necessarily following particular recommendations, as meticulous implementation of provenance can easily evolve into a substantial project of its own. This is why FREYA partners have deliberately concentrated on the implementation of certain aspects of provenance on the PID suppliers' end, which is being reported in D2.2, and on the discipline-specific notions of provenance presented by the project partners, which is the focus of this deliverable.

The main focus of this deliverable is the different approaches to provenance, as understood, expressed and implemented by the FREYA disciplinary partners in their various pilot applications. These presentations outline general approaches to provenance in the particular research context of each organisation, current

[1] An overview of the PROV Family of Documents: https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/
[2] RDA Research Data Provenance Interest Group: https://rd-alliance.org/groups/research-data-provenance.html
[3] RDA Provenance Patterns Working Group: https://www.rd-alliance.org/groups/provenance-patterns-wg

and future implementations, and describe provenance activities that are supported by persistent identifiers.

This deliverable follows the initial ambition and description of the pilot applications in Deliverable 4.1[4], and shows the progress that has been made in terms of provenance considerations in these communities. We hope the discourse within and across the pilot applications will be useful for triggering similar discussions within EOSC and in the wider Open Science communities. Updates on the work on provenance can be expected in the following deliverables of FREYA in this Work Package.

[4] FREYA Deliverable D4.1: https://zenodo.org/record/2414839

2 British Library

2.1 General approach to provenance

Provenance is defined at The British Library (BL) as information relating to the origin, source and curation of its collection items both digital and physical. It can also pertain to the source of metadata and documentation about the object.

Within BL collections, particularly heritage collections, provenance information is not always expressed as metadata within the resource but could be captured as documentation about the resource or even information contained within the resource itself e.g. stamps on a book. For archive collections, detailed provenance and custodial history information is described in the cataloguing of the resource to ISAD-G standard[5].

For the BL, provenance information is generally expressed through metadata, but not necessarily public-facing metadata and on occasion it may be included within a documentation file rather than within structured metadata.

The BL is currently interested particularly in trying to capture better provenance information about digital resources themselves and would like to use persistent identifiers where possible to support this, particularly where it is possible to provide metadata enriched with PIDs to enhance the understanding of the resource. Generally, the provenance of the resource is of greater interest than the provenance of the metadata.

2.2 Current provenance activities and implementations

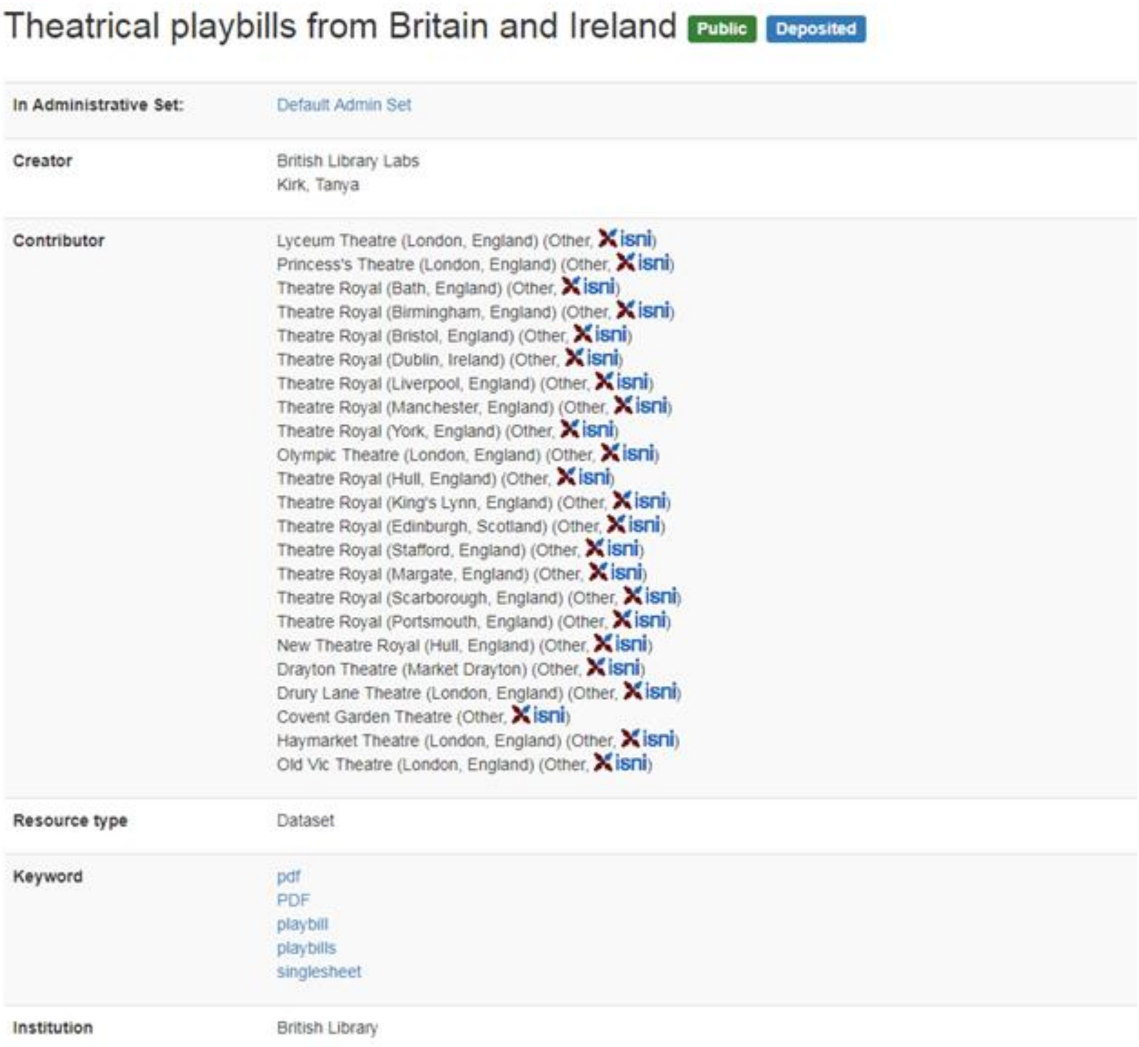
As part of the re-platforming of data.bl.uk, we have investigated the potential of extracting the creators of the individual files within large datasets to be included within the metadata of the resource to give richer provenance information. As a pilot implementation, persistent identifiers relating to contents of two datasets held on data.bl.uk have been added to the record to provide provenance information. However, the DataCite 4.1 metadata schema has meant we have had to make some choices about how the metadata is expressed which are not ideal.

For example, initially it was hoped to include ISNIs for the theatres, which have playbills included in the Theatrical Playbills digitised datasets, as Related Identifiers. However, this is not currently permitted within the DataCite schema, so it was included as a Contributor instead (Figure 1). Related Identifiers would be more appropriate given that although they are the origin of the original hard-copy playbills, they did not directly contribute to the dataset itself.

The Archive Resource Key (ARK)[6] identifiers that the Library routinely assigns to the digital objects were also added as Related Identifiers to the records, however Related Identifiers are currently not actionable links within the repository.

[5] ICA General International Standard Archival Description: https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

[6] Archive Resource Keys are a persistent identifier used for both digital and physical objects: http://n2t.net/e/ark_ids.html



Theatrical playbills from Britain and Ireland Public Deposited

In Administrative Set:	Default Admin Set
Creator	British Library Labs Kirk, Tanya
Contributor	Lyceum Theatre (London, England) (Other, isni) Princess's Theatre (London, England) (Other, isni) Theatre Royal (Bath, England) (Other, isni) Theatre Royal (Birmingham, England) (Other, isni) Theatre Royal (Bristol, England) (Other, isni) Theatre Royal (Dublin, Ireland) (Other, isni) Theatre Royal (Liverpool, England) (Other, isni) Theatre Royal (Manchester, England) (Other, isni) Theatre Royal (York, England) (Other, isni) Olympic Theatre (London, England) (Other, isni) Theatre Royal (Hull, England) (Other, isni) Theatre Royal (King's Lynn, England) (Other, isni) Theatre Royal (Edinburgh, Scotland) (Other, isni) Theatre Royal (Stafford, England) (Other, isni) Theatre Royal (Margate, England) (Other, isni) Theatre Royal (Scarborough, England) (Other, isni) Theatre Royal (Portsmouth, England) (Other, isni) New Theatre Royal (Hull, England) (Other, isni) Drayton Theatre (Market Drayton) (Other, isni) Drury Lane Theatre (London, England) (Other, isni) Covent Garden Theatre (Other, isni) Haymarket Theatre (London, England) (Other, isni) Old Vic Theatre (London, England) (Other, isni)
Resource type	Dataset
Keyword	pdf PDF playbill playbills singlesheet
Institution	British Library

Figure 1: A screenshot of an example record from the data.bl.uk collection populated with contributors reflecting the provenance of the resource. This record demonstrates some of the issues encountered with expressing provenance information through PID metadata: it is not possible to add ISNIs as related identifiers, so they were expressed as contributors instead.

2.3 Plans and considerations for the future

As the repository platform where data.bl.uk resides is developed the records will be marked up to comply with schema.org. It is also hoped to make further provenance metadata such as that described by the new DataCite activities API[7] visible on the repository landing pages. This is the only foreseen application of PROV within the service.

For new datasets being added to data.bl.uk, we will capture metadata to enhance the provenance information at the point of deposit as much as possible. We will also explore the possibility of developing innovative methods of displaying these records, which include large numbers (potentially in the thousands) of related identifiers and therefore do not lend themselves to easy display and could potentially hinder an end user's use of the resource. There are also plans to enhance the display of repository records to ensure that the PID Graph for these resources is completely actionable.

The BL's Digital Library System which manages all digitised and born-digital content uses the PREMIS metadata schema[8] that includes detailed provenance information regarding the source of the digital

[7] Exposing DOI metadata provenance: https://doi.org/10.5438/wy92-xj57

[8] PREMIS metadata schema: https://www.loc.gov/standards/premis/

objects and within certain collection types, such as digitised archive material, further provenance metadata is also gathered, relating to the physical resource. ARKs are used as an identifier of resources held within the Digital Library System and the provenance metadata is related to the ARK.

The Digital Library System is due to be replaced by a new system, due to go live after FREYA ends, for which preservation requirements were specified in detail, including provenance information. However, it is anticipated that there will be some work done in adapting the workflow of this new preservation system to capture this information.

2.4 Summary table

Definition	Resource provenance generally expressed as metadata but also as supporting documentation. Metadata provenance is of interest but secondary to resource provenance.
Purpose	The main purpose is in making the context of the resource more understandable. Enhancing the provenance information available also improves the audit information available for preservation purposes. One of the user stories gathered, stating a desire to link bespoke software available in GitHub used for analysis of research datasets is also relevant[9].
Method	Embedding provenance information in the metadata and enhancing the number of PIDs within the metadata to improve authority of the record.
Standards (current)	PREMIS is used for material within the Digital Library System which captures provenance information. Archive material is described using ISAD-G which includes the custodial history of the resource. No formal provenance standards are in use at present.
Standard integration (future)	It is planned to include schema.org mark up within the schedule of developments of the new repository. However, as the systems we are using are under development we are open to the possibility of incorporating PROV terminology such as that used by the DataCite activities APIs and to display that on landing pages. It is not anticipated to use any PROV terminology beyond that offered by DataCite as metadata provenance is of secondary interest to resource provenance.
Implications for the PID Graph	The plan to enhance metadata where possible with additional PIDs and capture it for new datasets creates a PID Graph for these resources and provides detailed authoritative provenance information. The use of PIDs which are centrally maintained to provide this provenance information enhances the authority of the information.

[9] FREYA user story on software linking: https://github.com/datacite/freya/issues/39

3 CERN

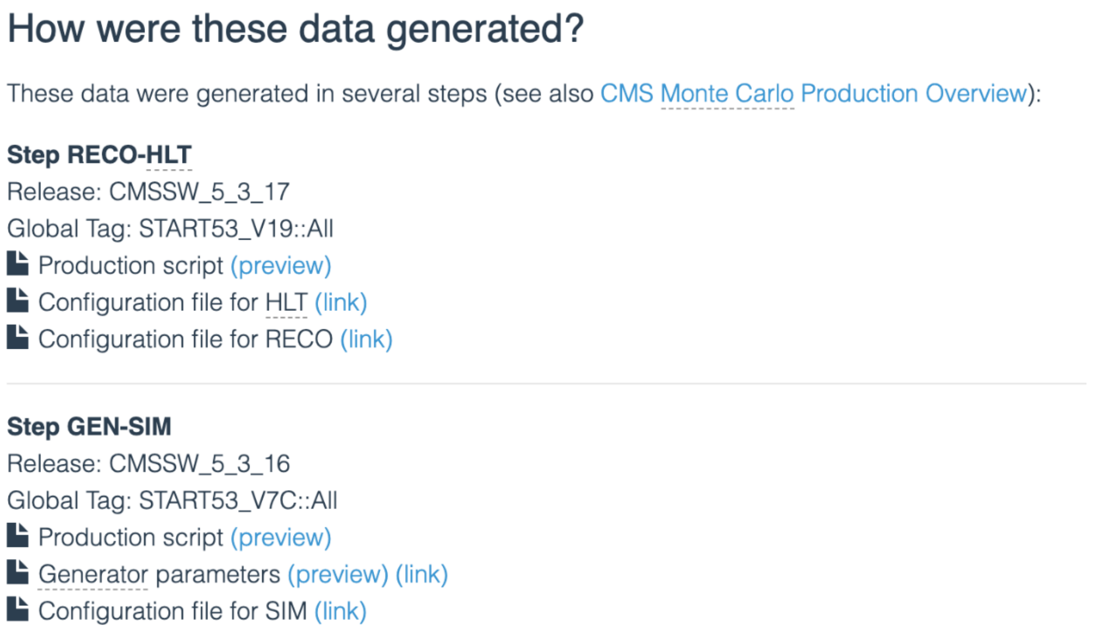
3.1 General approach to provenance

At CERN, various services serve the High-Energy Physics (HEP) and adjacent communities. It is therefore challenging to conclude that there is a common understanding and approach to issues like provenance. However, focusing on the core CERN services that use PIDs and deal with scholarly information and digital objects, provenance currently refers to information that provides insight on how a resource came to be, i.e. resource provenance. The term also refers to the provenance of the metadata itself.

3.2 Current provenance activities and implementations

CERN's main pilot applications in FREYA where the topic of provenance is most relevant are the CERN Open Data Portal (COD)[10], an open access data repository, and CERN Analysis Preservation (CAP)[11], a data preservation service (access-restricted tool). CERN's main provenance use case is about resource provenance. As we are working very closely with the community to develop these services, we are very aware of their needs and contextual information at a granular level is always important. Provenance information is part of that contextual information and in most cases the most useful provenance information for the community is details on how a resource was generated (all the methodology and processing steps, software used, other related datasets, etc.).

In COD, users benefit from customised manually curated metadata. A lot of effort goes into generating rich resource provenance information for the records (Figure 2).



How were these data generated?

These data were generated in several steps (see also CMS Monte Carlo Production Overview):

Step RECO-HLT
Release: CMSSW_5_3_17
Global Tag: START53_V19::All
Production script (preview)
Configuration file for HLT (link)
Configuration file for RECO (link)

Step GEN-SIM
Release: CMSSW_5_3_16
Global Tag: START53_V7C::All
Production script (preview)
Generator parameters (preview) (link)
Configuration file for SIM (link)

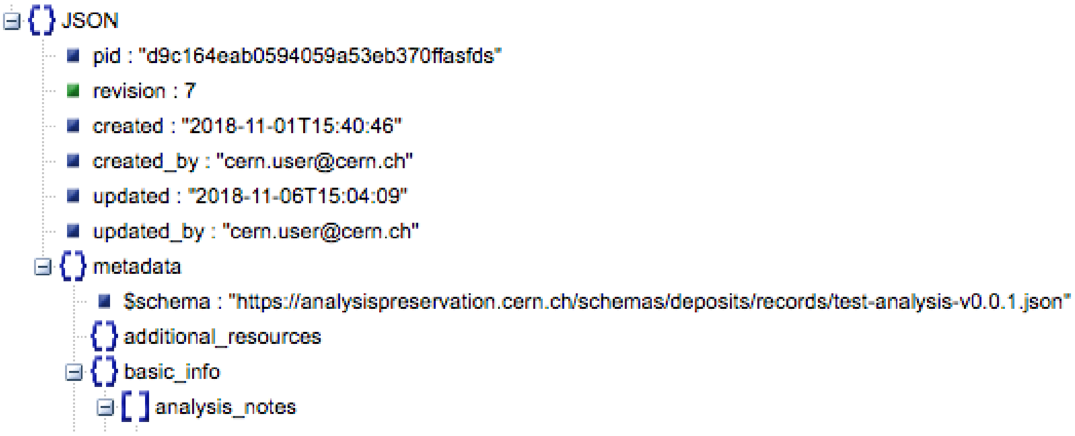
Figure 2: Part of a CERN Open Data bibliographic record showing the processing steps for the generation of a specific dataset.

Information on provenance creates trust in the resources shared openly and facilitates reuse by helping users understand the full context of a published resource. Enabling reuse is one of the main goals of COD, which makes this provenance use case very important. Since GitHub is used for content management for

[10] CERN Open Data portal: http://opendata.cern.ch/
[11] CERN Analysis Preservation (GitHub): https://github.com/cernanalysispreservation/analysispreservation.cern.ch

this service, it is possible to track all metadata changes. However, this is not a very accurate way of capturing provenance, because information is often inputted through scripts.

CAP provides rich metadata that focuses on the comprehensive coverage of a physics analysis and all its components (resource provenance). Provenance information is needed to enable reuse through better understanding about “who did what and when”. Users are able to input information themselves, which makes capturing metadata provenance crucial. Currently, we are able to get some information about changes made to analyses. Figure 3 shows part of the metadata of an example analysis in CAP; from there, it is possible to determine who created and/or updated an analysis, when, and how many revisions there are.



```
{ } JSON
    pid : "d9c164eab0594059a53eb370ffasfds"
    revision : 7
    created : "2018-11-01T15:40:46"
    created_by : "cern.user@cern.ch"
    updated : "2018-11-06T15:04:09"
    updated_by : "cern.user@cern.ch"
  { } metadata
      $schema : "https://analysispreservation.cern.ch/schemas/deposits/records/test-analysis-v0.0.1.json"
    { } additional_resources
    { } basic_info
      [ ] analysis_notes
```

Figure 3: Metadata provenance captured about an analysis in CERN Analysis Preservation.

Generally, services like CAP and COD aim at providing high-quality information about resources as a means to enable reuse of research materials. The more contextual information, the more useful a resource is to the community or the external users in the case of public-facing services.

In terms of standards, for COD, the data model has taken elements from several standards: DCAT, Dublin Core and DataCite Metadata Schema in order to capture the more high-level information and uses customised metadata fields for discipline-specific information. In addition, other standards such as ISO 8601 and ISO 639-2 (for dates and language names respectively) are also used to ensure that as much information as possible is expressed in a standardised way. In CAP, the schemata are customised to each collaboration and are very granular and detailed as they need to capture all elements of physics analyses. No dedicated provenance standard is used because, as mentioned above, the most prominent provenance use case so far has been showcasing human-readable provenance information in the user interface.

Finally, related to the PID Graph topic, using and connecting to services that collect metadata on PIDs (e.g. DataCite or Crossref) means that we can have access to much more provenance information than what we ourselves expose.

3.3 Plans and considerations for the future

The presented CERN services are user-driven in their development and prioritisation of features. The services are developed in close collaboration with representatives from the experimental collaborations at CERN who are in a position to give input and express the needs of their communities. So far there has not been a use case from the overall HEP community to use PROV; the metadata we create or harvest at present reflect the needs of the community and the collaborations of the large-scale LHC experiments.

However, it should be noted that we intend to focus on provenance more in the future. DataCite’s recent implementation of PROV resonates in the community and HEP’s public/open services could consider a similar implementation.

Furthermore, schema.org markup using JSON-LD has been implemented (see D4.1) for a basic set of metadata in COD and we intend to extend it in the future. This implementation has helped improving the discoverability of the research resources through search engines, like Google Dataset Search[12]. Future extension of schema.org for COD could include provenance metadata, but this will need to be decided based on community needs. For CAP, we are also exploring the option of implementing CodeMeta[13], which focuses on metadata for software, but this is still in the research phase.

All the services mentioned above are using the open source software Invenio[14]. Metadata provenance is captured in the record change history. Taking advantage of that, we could generate a "live" metadata provenance log for the analyses that users input to improve the trust into the research objects preserved. For this use case, indeed the PROV nomenclature could be used, if considered beneficial for the community.

Other considerations relate to cases of tracking the history of the origin of the metadata. For example, INSPIRE[15] is the core HEP information system that aggregates content from multiple sources. The challenge here is dealing with metadata from various sources during the curation process: metadata from the original place of publication, metadata created by the aggregator platform itself and so on. It is often the case that (some) metadata for a resource is taken by publishing services from where it was first published. In the case of INSPIRE, records are manually curated; something to consider would be generating and exporting detailed provenance metadata to keep track of all these actions in an easier way.

Finally, the use of new or emerging PID types will certainly enhance the quality of provenance information as PIDs further facilitate discoverability. Integrating more PIDs, e.g. instrument PIDs, will be an important part of resource provenance implementations in the future and will further extend CERN's PID graph.

3.4 Summary table

Definition	Provenance refers to information included in the metadata of a digital object that provides insight on how it came to be. The term also refers to the provenance of the metadata itself.
Purpose	Resource provenance is crucial for reuse and context. Resource provenance has proven helpful for building trust into the research objects on COD and CAP as it helps understanding who did what, when, how.
Method	User-driven metadata enrichment: provenance information included in metadata, understanding what kind and what level of metadata granularity the community needs.
Standards (current)	Though no dedicated provenance standards are used, some metadata and resource provenance information are captured.
Standard integration (future)	Schema.org enrichment and perhaps API enrichment corresponding to the work of DataCite on provenance. Considerations of generating metadata provenance from record change history information using PROV. Possible integration of CodeMeta integration in CAP.

[12] Google Dataset Search: https://toolbox.google.com/datasetsearch
[13] The CodeMeta Project: https://codemeta.github.io/index.html
[14] Invenio: https://invenio-software.org/
[15] INSPIRE: https://inspirehep.net

Implications for the PID Graph	Adding information on new PID types will enrich the PID graph (e.g. a dataset was generated by a certain piece of equipment). Using services that collect metadata on PIDs (e.g. DataCite or Crossref) means that we can have access to much more provenance information than what we ourselves expose.

4 DANS

4.1 General approach to provenance

At DANS, provenance is information about the origin and source of (meta)data, but also the history of ownership and information about the curation of data. For archival purposes provenance needs to be a backlog for all steps taken during data curation.

4.2 Current provenance activities and implementations

NARCIS is a national portal which aggregates metadata from all Dutch scientific research institutes. It provides information about publications, datasets, research projects, researchers and research organisations.

NARCIS provides OAI-PMH[16] provenance information for all the records it re-publishes. It holds information about the originating repository, including original identifier, source URL and harvest date. PIDs also contain information about provenance:

- At a national level: all organisations participating in the Dutch National Infrastructure assign a URN:NBN to every digital object with a prefix. A URN:NBN is a “National Bibliographic Number” and a resolvable PID. The prefix represents the organisation and it is an important part of the provenance of an object. The Royal Library of the Netherlands plays an important role in long-term preservation of all Open Access publications within the Dutch infrastructure and the prefix gives information about the origin of an object.
- At the repository level: some universities and other research organisations assign a handle and other provenance data about the origin of an object.

EASY is a Trusted Digital Repository and archive for the Social Sciences and Humanities. In the EASY archive, provenance implementations include the following:

- In the administrative metadata, all actions of DANS data managers are logged. Each curation step of the dataset is traceable up to the deposit.
- All original files are stored in a separate folder and are immutable. For dissemination purposes files can be downloaded from another folder and files are, when needed, converted for dissemination purposes.
- DANS keeps provenance metadata from all steps in the curation process: from deposit to publishing and maintaining.
- During our dataset deposit process, we keep provenance metadata for all actions DANS carries out with the data, including checksum, virus checks, etc.
- DANS sends to the depositor a deposit agreement with a list of all files (including checksum), a PID, and other information. The depositor can always check if the original files are still in the archive.

4.3 Plans and considerations for the future

The DANS EASY team is discussing the implementation of W3C PROV standards in the near future. We believe the vocabulary of PROV could help us address specific use cases, e.g. PROV offers “labels” to identify the person or organisation who has to pay for the storage of data, or who owns the data. In addition to the benefits of the rich vocabulary, DANS believes it is important to use an international standard, especially for data exchange and compatibility with other services.

[16] Open Archives Initiative Protocol for Metadata Harvesting: https://www.openarchives.org/pmh/

4.4 Summary table

Definition	Provenance is information about the origin and source of (meta)data, but also the history of ownership and information about curation of data.
Purpose	• EASY archive: information about origin, history of every step in the curation of research data. • NARCIS metadata aggregator: mainly to inform the user about the origin of a metadata record and location of the digital object.
Method	• EASY Archive: every archival step is logged and part of the metadata. Original files are always kept separately from dissemination copies. • NARCIS: OAI-PMH provenance information.
Standards (current)	Special fields in internal XML format and OAI-PMH provenance information tags[17] within each metadata record.
Standard integration (future)	• EASY archive: currently DANS is rebuilding the archive. In the near future PROV could be a valuable vocabulary to describe provenance information. • NARCIS: NARCIS currently supports OAI-PMH provenance. No direct need to use other standards at this moment. For the future, PROV could be used in our JSON for Linked Data.
Implications for the PID Graph	PROV in schema.org or JSON-LD could contain information about the timestamp and origin of PIDs. A relevant use case could be about relations between publisher PIDs (article) and repository PIDs (preprint, or copy of same article, including timestamps).

[17] Schema for the description of the provenance of metadata that is re-exposed by an OAI repository: http://www.openarchives.org/OAI/2.0/provenance.xsd

5 EMBL-EBI

5.1 General approach to provenance

EMBL-EBI, "home to big data in biology", considers the provenance of a dataset to be the information defining the "source" of the data. Thus, a sample is considered to be the provenance: it is the origin/source of any resulting data, be it a direct property of the sample, e.g. nucleotide sequence, or a result after using it as a component of further research.

The EMBL-EBI hosts numerous life science data resources[18] and importantly makes them freely and openly available to users. The data resources include deposition databases (that index primary data types such as samples, nucleotide sequences, protein structures or research papers) and knowledge-bases, where the primary data type is curated information, e.g. enzyme catalysed reactions. Some of the data resources at EMBL-EBI serve as both deposition and knowledgebase archives.

Deposition databases have associated metadata that indicates the provenance. These are records that indicate how the sample/data was acquired, by whom, when, and from where. For the latter, the sample is key, as is the granularity: the provenance for a protein sequence would include information about the protein, as well as the organism, organ/tissue/cell or cell line from which the sequence originates.

In the case of knowledge-bases, the provenance of a record (source of information) can be made up of datasets deposited in another database. There is a provenance chain built on trust in these databases. A complex cascade can be seen in Figure 4 below.

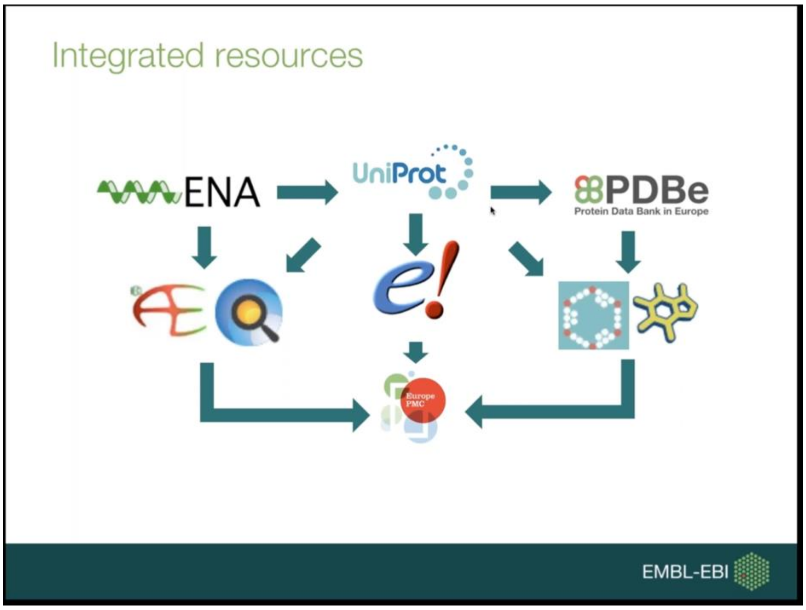


Figure 4[19]: Data provenance in life sciences - a cascade built on trusted resources. This schematic represents how data resources at EMBL-EBI are integrated. A literary record in Europe PMC (bottom) may mention an expression record from ArrayExpress Expression Atlas, or genome information from Ensembl (e!) or a chemical compound from ChEMBL or ChEBI (middle), that draws on a protein structure record in PDBe. These in turn draw protein sequence information initially from UniProt, and/or primary sequence information from the ENA. The ENA record would be deemed sufficient provenance for a nucleotide sequence mentioned in a curated record in UniProt, ArrayExpress or Europe PMC, for example. Including mention of the accession number for the ENA record in a subsequent curated record is essential for provenance; including the accession number as an actionable link within the curated record would be ideal.

[18] EMBL-EBI Tools & Data Resources: https://www.ebi.ac.uk/services

[19] Figure courtesy of Introduction to EMBL-EBI resources, EBI Training Online: https://www.ebi.ac.uk/training/online/course/introduction-embl-ebi-resources-webinar

ELIXIR, the European Research Infrastructure for life sciences data[20], brings together life sciences resources from across Europe and has a key influence on data management in European life sciences. ELIXIR's interoperability platform[21] aims to "help people and machines to discover, access, integrate and analyse biological data. It encourages the life science community to adopt standardised file formats, metadata, vocabularies and identifiers". This includes dealing with provenance tracking, encouraging provenance standards and providing useful resources to this end. EMBL-EBI is a member of ELIXIR, and Europe PMC, hosted by EMBL-EBI, is recognised as one of ELIXIR's 20 core data resources, i.e. considered to be of most fundamental importance to the life science community for the long-term preservation of biological data[22].

5.2 Current provenance activities and implementations

Europe PMC's primary entry type is literature, most commonly journal article publications. At Europe PMC journal article content is extensively linked to data resources and ORCID iDs, and funding information (i.e. funders and grants) is captured. This rich interconnected knowledge base is provided to users via programmatic access and search tools on a user interface.

Data referencing within research articles and preprints is weak in the life sciences: where authors do mention specific data sources this is often without including actionable links. To enrich the research provenance information provided in a literary article, Europe PMC offers the following:

- Additional information about data resources used in a literary article, including:
    - External data resources that point to the Europe PMC article - links are provided for these allowing the reader to move between the literary record and the data resource record.
    - Text-mined accession numbers (i.e. persistent identifiers of datasets) - text mined terms are highlighted for users and accompanied by a pop-up box providing: a link to the related database record where available, the source of the annotation, and a feedback link where readers can verify whether the information is useful or incorrect[23] (see Figure 5). Note that text mining results are displayed on full text records only if published using a license permitting reuse.
- Links to grants where the information is provided by the authors, or present within the grants database of Europe PMC funders[24].
- Links to other relevant external information provided by third-party data miners - these are referred to as "external links"[25].
- Search functionality to identify publications that have Data Availability Statements[26].

Data provenance information is collected and provided to users to assure them of the rigor of the dataset (the context in which the data was collected and the relationship to other datasets) and to enable reproducibility.

[20] ELIXIR Europe: www.elixir-europe.org

[21] ELIXIR Europe Interoperability Platform: https://elixir-europe.org/platforms/interoperability

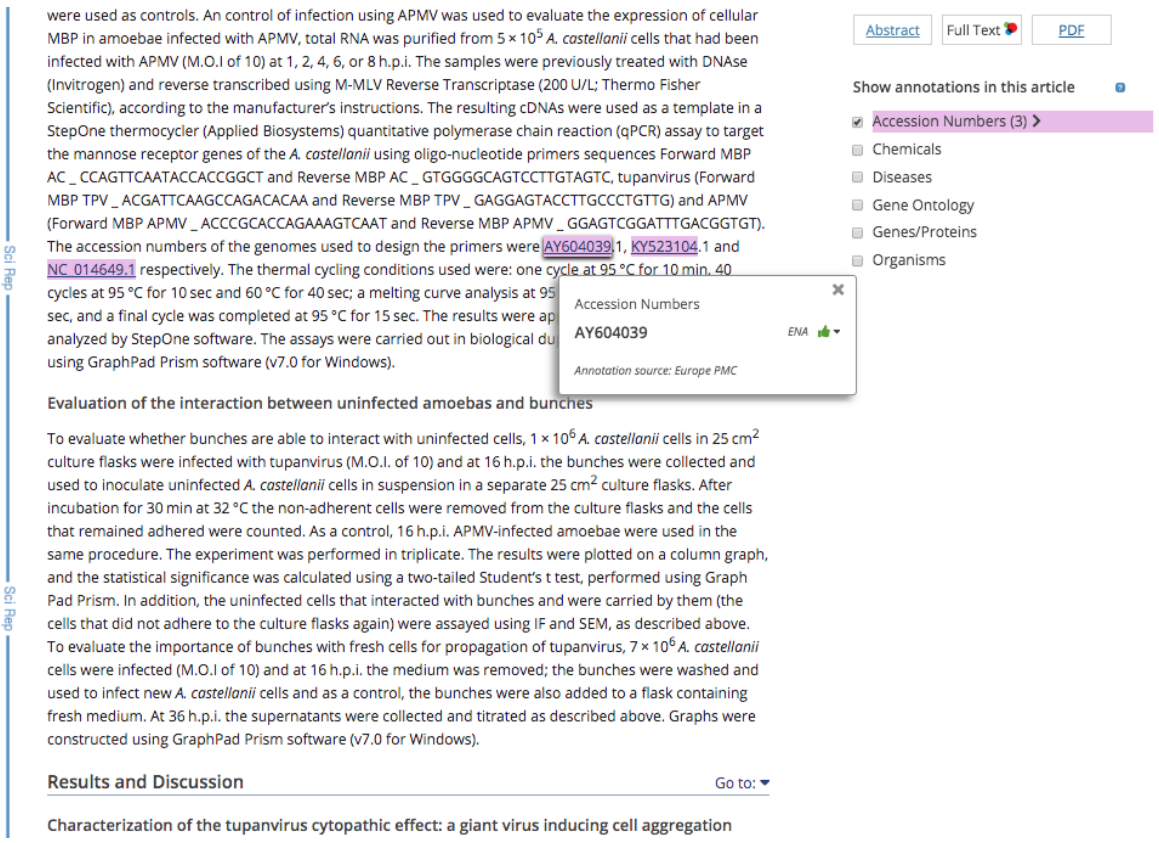
[22] The ELIXIR core data resources: fundamental infrastructure for the life sciences (preprint): https://www.biorxiv.org/content/10.1101/598318v1

[23] How to use SciLite Annotations in Europe PMC: https://europepmc.org/Annotations#how-to-use-sci-annot

[24] Europe PMC Grant Finder: https://europepmc.org/grantfinder

[25] How can I find external links in Europe PMC: https://europepmc.org/Help#findExternalLinks (see tab in Figure 6)

[26] The Data Availability Statement is a section of a literary record that contains guidelines on data access: it states whether the data is available, underlines conditions for access, and includes hyperlinks to publicly archived datasets analysed or generated during the study. To promote reproducibility of results and reuse of datasets, an increasing number of journals require inclusion of these statements in the articles they publish. Europe PMC monitors these FAIR efforts and provides the functionality to search for literary records that contain Data Availability Statements. Currently over 295.000 full-text publications in Europe PMC contain a dedicated data availability section. See: https://europepmc.org/search?query=DATA_AVAILABILITY%3A*



were used as controls. An control of infection using APMV was used to evaluate the expression of cellular MBP in amoebae infected with APMV, total RNA was purified from 5 × 10^5 *A. castellanii* cells that had been infected with APMV (M.O.I of 10) at 1, 2, 4, 6, or 8 h.p.i. The samples were previously treated with DNAse (Invitrogen) and reverse transcribed using M-MLV Reverse Transcriptase (200 U/L; Thermo Fisher Scientific), according to the manufacturer's instructions. The resulting cDNAs were used as a template in a StepOne thermocycler (Applied Biosystems) quantitative polymerase chain reaction (qPCR) assay to target the mannose receptor genes of the *A. castellanii* using oligo-nucleotide primers sequences Forward MBP AC _ CCAGTTCAATACCACCGGCT and Reverse MBP AC _ GTGGGGCAGTCCTTGTAGTC, tupanvirus (Forward MBP TPV _ ACGATTCAAGCCAGACACAA and Reverse MBP TPV _ GAGGAGTACCTTGCCCTGTTG) and APMV (Forward MBP APMV _ ACCCGCACCAGAAAGTCAAT and Reverse MBP APMV _ GGAGTCGGATTTGACGGTGT). The accession numbers of the genomes used to design the primers were AY604039.1, KY523104.1 and NC_014649.1 respectively. The thermal cycling conditions used were: one cycle at 95 °C for 10 min, 40 cycles at 95 °C for 10 sec and 60 °C for 40 sec; a melting curve analysis at 95 sec, and a final cycle was completed at 95 °C for 15 sec. The results were ap analyzed by StepOne software. The assays were carried out in biological du using GraphPad Prism software (v7.0 for Windows).

Evaluation of the interaction between uninfected amoebas and bunches

To evaluate whether bunches are able to interact with uninfected cells, 1 × 10^6 *A. castellanii* cells in 25 cm^2 culture flasks were infected with tupanvirus (M.O.I. of 10) and at 16 h.p.i. the bunches were collected and used to inoculate uninfected *A. castellanii* cells in suspension in a separate 25 cm^2 culture flasks. After incubation for 30 min at 32 °C the non-adherent cells were removed from the culture flasks and the cells that remained adhered were counted. As a control, 16 h.p.i. APMV-infected amoebae were used in the same procedure. The experiment was performed in triplicate. The results were plotted on a column graph, and the statistical significance was calculated using a two-tailed Student's t test, performed using Graph Pad Prism. In addition, the uninfected cells that interacted with bunches and were carried by them (the cells that did not adhere to the culture flasks again) were assayed using IF and SEM, as described above. To evaluate the importance of bunches with fresh cells for propagation of tupanvirus, 7 × 10^6 *A. castellanii* cells were infected (M.O.I of 10) and at 16 h.p.i. the medium was removed; the bunches were washed and used to infect new *A. castellanii* cells and as a control, the bunches were also added to a flask containing fresh medium. At 36 h.p.i. the supernatants were collected and titrated as described above. Graphs were constructed using GraphPad Prism software (v7.0 for Windows).

Results and Discussion

Characterization of the tupanvirus cytopathic effect: a giant virus inducing cell aggregation

Figure 5: Provenance information for a term text-mined in Europe PMC for article DOI: 10.1038/s41598-018-36552-4. Text mining by Europe PMC after publication picks up 3 accession numbers - shown highlighted in purple. Hovering over a highlighted term, provides the provenance information: in this case it is Accession number AY604039; source of text mining is Europe PMC; the accession number has been verified by the ENA, as an accession number for a nucleotide sequence[27].

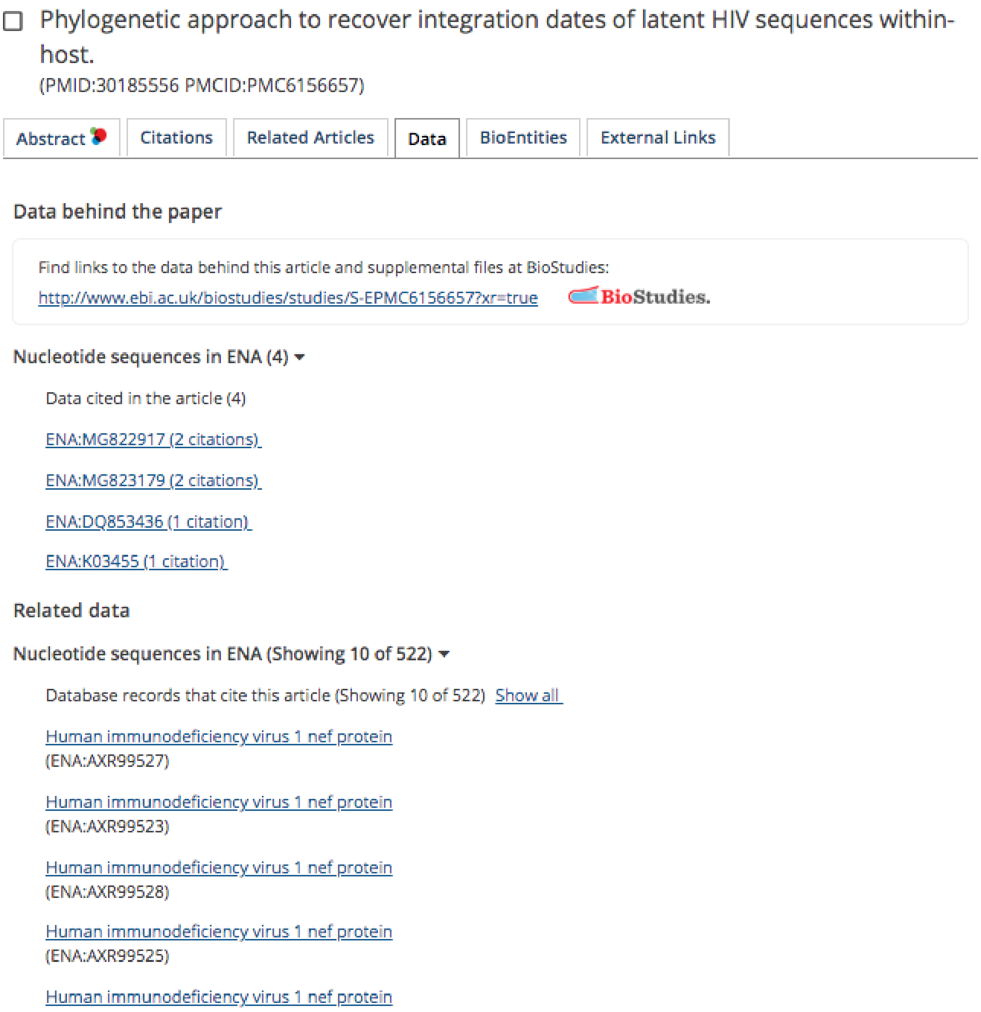
Standards used at Europe PMC for capturing provenance and metadata: literary records are encoded in XML JATS (Journal Article Tag Suite, version 1.2 (ANSI/NISO Z39.96-2019)). However, not all data references are encoded by users (i.e. publishers). To ensure a greater level of integration, Europe PMC captures data references as external links and annotations using the following standard formats:

- The datalinks are represented in the Scholix format[28] and made available to the end user both via APIs and in the data tab of a literary record. The data tab (see Figure 6) comprises a mash up of external references from external data resources that point to the publication, as well as Europe PMC's text-mined accession numbers.
- Europe PMC uses the W3C Web Annotation Data Model[29], in which the creator of the record is a required field. This ensures that the end user can identify who provided the annotated information, thus increasing transparency and trust. An RDF representation was considered originally for annotated datasets within Europe PMC, but the alternative JSON-LD was favoured going forward.

None of the EMBL-EBI resources utilise PROV-O. An RDF platform was initiated to support several data resources at EMBL-EBI[30]. Some PROV terms are being used in the RDF platform but not systematically.

Europe PMC also uses schema.org. Bioschema.org is an extension of schema.org adopted for the life sciences. ELIXIR, via its interoperability platform, is actively promoting the use of Bioschema.org to make life science data resources across Europe more interoperable and findable.

[27] The corresponding ENA record is: https://www.ebi.ac.uk/ena/data/view/AY604039
[28] About Scholix: http://www.scholix.org/about
[29] W3C Web Annotation Data Model: https://www.w3.org/TR/annotation-model/
[30] EMBL-EBI RDF platform: https://www.ebi.ac.uk/rdf/



☐ Phylogenetic approach to recover integration dates of latent HIV sequences within-host.
(PMID:30185556 PMCID:PMC6156657)

Data behind the paper

Find links to the data behind this article and supplemental files at BioStudies:
http://www.ebi.ac.uk/biostudies/studies/S-EPMC6156657?xr=true

Nucleotide sequences in ENA (4)

Data cited in the article (4)

ENA:MG822917 (2 citations)

ENA:MG823179 (2 citations)

ENA:DQ853436 (1 citation)

ENA:K03455 (1 citation)

Related data

Nucleotide sequences in ENA (Showing 10 of 522)

Database records that cite this article (Showing 10 of 522) Show all

Human immunodeficiency virus 1 nef protein
(ENA:AXR99527)

Human immunodeficiency virus 1 nef protein
(ENA:AXR99523)

Human immunodeficiency virus 1 nef protein
(ENA:AXR99528)

Human immunodeficiency virus 1 nef protein
(ENA:AXR99525)

Human immunodeficiency virus 1 nef protein

Figure 6: The data tab for a literary record (DOI: 10.1073/pnas.1802028115) in Europe PMC. The tab records links to provenance information for data resources used within the study. The data tab reveals that this Europe PMC record has an associated BioStudies record that was generated after publication and indexing of the article. Note the separate tabs where BioEntities and External Links are displayed.

5.3 Plans and considerations for the future

Given the life science view that provenance is considered to be the information defining the "source" of the data, future work includes the following:

- In support of the FAIR principles, Europe PMC continues to promote use of the BioStudies database[31] to collect all the data behind a publication: linking to core life science databases, generic databases and storing supplemental files where necessary.
- Indexing preprints from a growing number of preprint servers and linking them to subsequently peer-reviewed publication (as well as to other versions of the preprints themselves)[32].
- Increasing the scope of text mining internally and advocacy to involve third parties to enrich information relevant to literary records such as funding information, performance metrics, links to curation efforts that support the findings in papers, e.g. instances where models have been actively verified.

[31] Sarkans et al. The BioStudies database - one stop shop for all data supporting a life sciences study, Nucleic Acids Research, 2017. DOI: https://doi.org/10.1093/nar/gkx965

[32] See FREYA Deliverable 4.3 for more details about Europe PMC's efforts to expose preprint versioning for users.

As mentioned above, EMBL-EBI is a member of the ELIXIR infrastructure and is also home to ELIXIR's administrative hub. ELIXIR operates nodes across 23 countries in Europe[33]. ELIXIR provides a forum for networking where data resources hosted across Europe can operate as a community to consider data management strategies and provenance standards for the life sciences.

5.4 Summary table

Definition	For EMBL-EBI we focus on resource provenance: the source of the data being recorded, who collected the resource/data and how it was collected.
Purpose	Endows trustworthiness; ensures rigor, reproducibility - can confer credit to the researcher responsible.
Method	Provenance of an annotation and the use of actionable identifiers - source is implicit from the identifier (e.g. PMCID is from PMC USA; PDB:4RQV is from the PDB).
Standards (current)	The following are used within Europe PMC: • XML (JATS: Journal Article Tag Suite, version 1.2 (ANSI/NISO Z39.96-2019)) • schema.org • RDF graph was considered but discontinued • scholix (e.g. in data tab for records) • web annotation data model (W3C standard); RDF used, now JSON-LD ELIXIR (via its Interoperability platform) strongly promotes the use of Bioschemas.
Standard integration (future)	PROV is not in use at EMBL-EBI. An RDF platform was initiated to support several data resources at EMBL-EBI. Some PROV terms are being used in the RDF platform, but not systematically. Other W3C standards are in use. The use case for implementing PROV in Europe PMC is unclear.
Implications for the PID Graph	PIDs form part of the provenance metadata. Providing this information addresses: transparency/confidence especially if metadata revealing the source of any data is exposed for the user of the search interface, e.g. for annotations derived by external text mining.

[33] About ELIXIR Europe: https://elixir-europe.org/about-us/who-we-are

6 PANGAEA

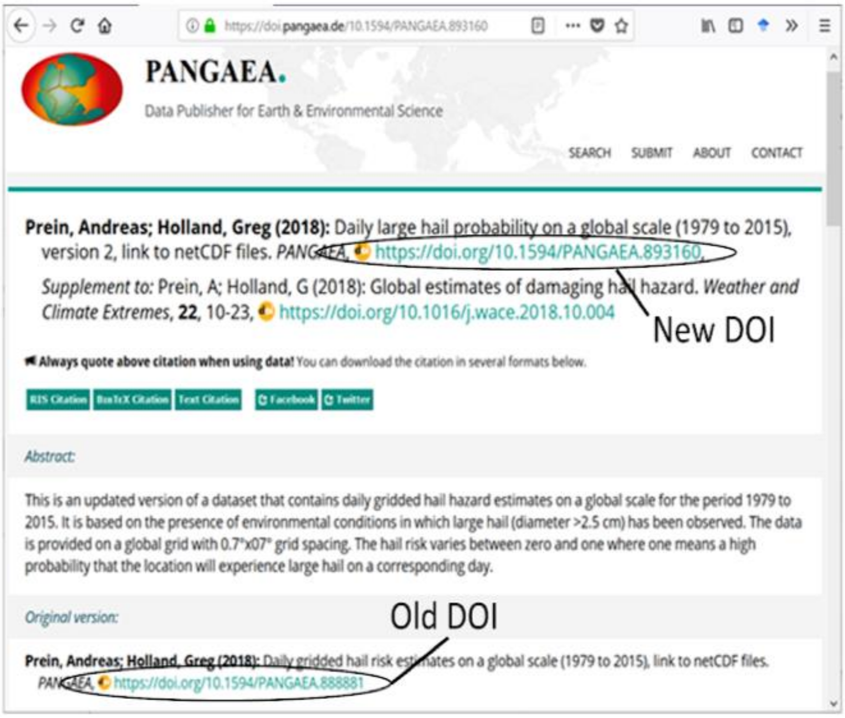
6.1 General approach to provenance

For PANGAEA, as a data publisher, the core definition of provenance is information that identifies the origin and history of a dataset, facilitating full transparency of changes through time. Versioning of datasets is a significant part of the provenance, keeping records of changes to a dataset, even past the initial DOI assignment. PANGAEA is working on expanding the amount of information provided as metadata, including metadata for instrumentation, preferably as PIDs. In a larger perspective, this is also part of a dataset's provenance, providing information about the procurement of the dataset. The PID Graph can be used to link the dataset with the specific instrumentation used for its procurement.

6.2 Current provenance activities and implementations

All datasets in PANGAEA are assigned a DOI and DOI metadata is published through DataCite. Provenance information about the metadata can be collected using the DataCite provenance API. This can be used for tracing the history of metadata versions. Recent developments, conducted as part of WP2, allow for the collection of more detailed information, including changes to the URL as well as identifying the specific metadata properties that were changed. This feature was initiated in March 2019 and henceforth changes to DOIs will be recorded automatically.

Versioning of datasets is an essential part of provenance, keeping record of changes to a dataset, even past the initial DOI assignment. Occasionally, published datasets within PANGAEA need to be altered from their original state. This usually happens per request of the authors wanting to update their dataset for corrections of common errors discovered at a later stage or implementation of quality control measures applied after the curation of the dataset. The history of these changes is recorded as dataset provenance information. The "original" dataset is archived and the "new" dataset is linked to its original source to ensure that the provenance information is not lost.



Prein, Andreas; Holland, Greg (2018): Daily large hail probability on a global scale (1979 to 2015), version 2, link to netCDF files. *PANGAEA*, https://doi.org/10.1594/PANGAEA.893160,

Supplement to: Prein, A; Holland, G (2018): Global estimates of damaging hail hazard. *Weather and Climate Extremes*, **22**, 10-23, https://doi.org/10.1016/j.wace.2018.10.004

Always quote above citation when using data! You can download the citation in several formats below.

Abstract:

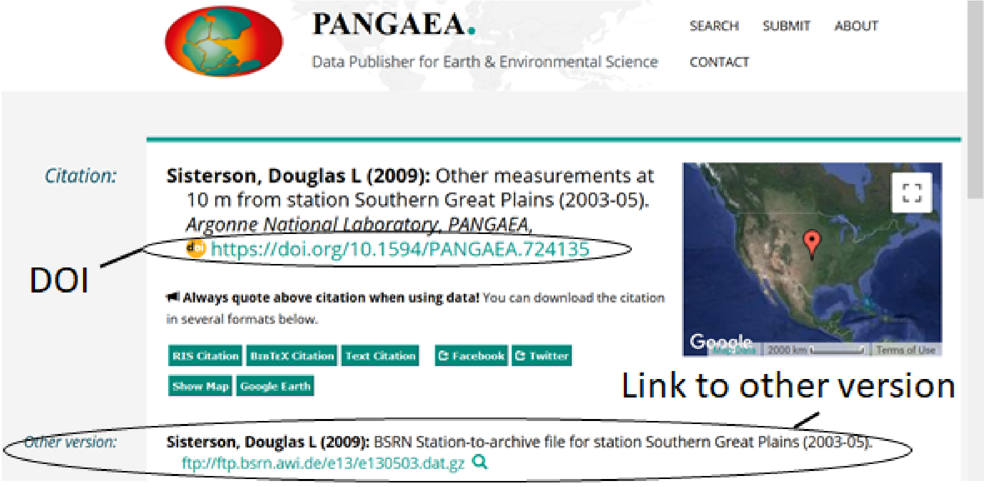
This is an updated version of a dataset that contains daily gridded hail hazard estimates on a global scale for the period 1979 to 2015. It is based on the presence of environmental conditions in which large hail (diameter >2.5 cm) has been observed. The data is provided on a global grid with 0.7°x07° grid spacing. The hail risk varies between zero and one where one means a high probability that the location will experience large hail on a corresponding day.

Original version:

Prein, Andreas; Holland, Greg (2018): Daily gridded hail risk estimates on a global scale (1979 to 2015), link to netCDF files. *PANGAEA*, https://doi.org/10.1594/PANGAEA.888881

Figure 7: Versioning of datasets in PANGAEA. When changes have been applied to the content of a published dataset, a new DOI is assigned and the PANGAEA database refers back to the original version.

PANGAEA has several ways of facilitating versioning of a dataset depending on the changes made. If a dataset is published, it cannot be changed anymore. Its integrity can be controlled by using a version identifier which is unique to the content of the dataset and which changes whenever the content is modified. Hence, if the content has changed, an updated DOI assignment is required (Figure 7). In cases where only format has changed but not the content, PANGAEA has several ways to refer to previous versions in different formats (Figure 8).



Citation: **Sisterson, Douglas L (2009):** Other measurements at 10 m from station Southern Great Plains (2003-05). *Argonne National Laboratory, PANGAEA,* https://doi.org/10.1594/PANGAEA.724135

Always quote above citation when using data! You can download the citation in several formats below.

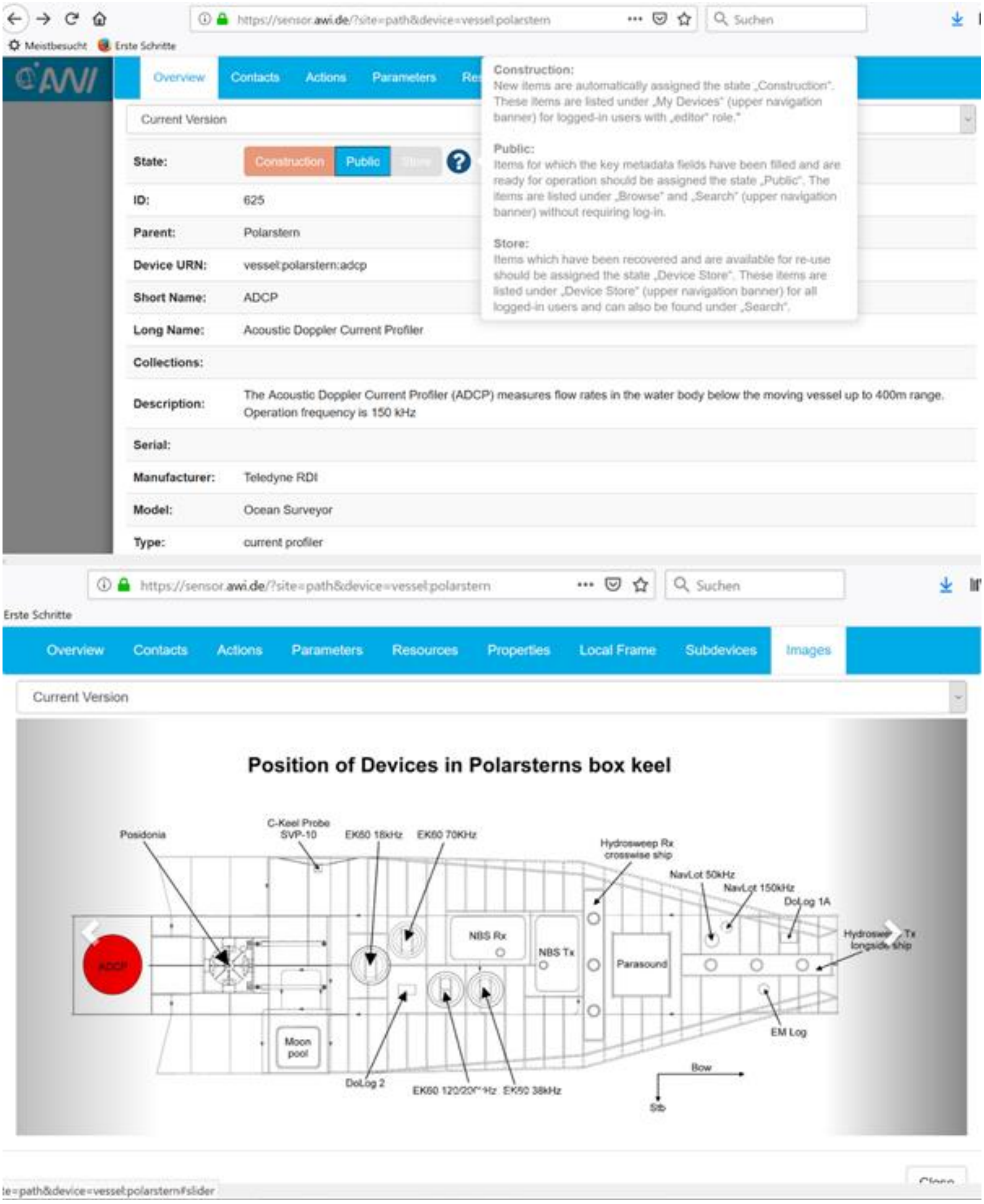
Other version: **Sisterson, Douglas L (2009):** BSRN Station-to-archive file for station Southern Great Plains (2003-05). ftp://ftp.bsrn.awi.de/e13/e130503.dat.gz

Figure 8: Versioning of datasets in PANGAEA. When changes have been applied to the format of a published dataset, the database shows a link to "other version".

6.3 Plans and considerations for the future

Future work on provenance for PANGAEA includes the implementation of identifiers for instruments. In collaboration with the Alfred Wegener Institute that recently has developed a sensor information system (Sensor.awi), PANGAEA will start to integrate PIDs for instruments at a prototype level. Sensor.awi is a registry for sensors used in marine research that facilitates access to and management of sensor metadata including the generation of identifiers for specific instruments. To enable management of the large and complex assortment of instruments, a set of essential core metadata is required for an instrument to be accepted in the registry.

The goal is to implement these identifiers for instrumentation in the metadata of datasets submitted to PANGAEA. This will allow the users of the PANGAEA database to specifically identify the instrument that has generated the dataset in question and get metadata about the instrument from Sensor.awi. This will be a valuable addition to the dataset's provenance. Sensor.awi registers important provenance information about the instruments and the platforms that carry them, such as ship, landers, moorings and remotely-operated vehicles. As part of the metadata, the Sensor.Awi registry provides both general and technical information about the instrument and its subdevices (Figure 9). In addition, important provenance information is provided about history of application, servicing, calibration and points of contact, as well as the specific location onboard research vessels or other platforms (Figure 9). Information like this can be very pertinent to the interpretation of data. Hence, this is important provenance information for the datasets generated from this instrument, which is not part of any metadata schema.



State: Construction | Public | Store

ID: 625

Parent: Polarstern

Device URN: vessel:polarstern:adcp

Short Name: ADCP

Long Name: Acoustic Doppler Current Profiler

Collections:

Description: The Acoustic Doppler Current Profiler (ADCP) measures flow rates in the water body below the moving vessel up to 400m range. Operation frequency is 150 kHz

Serial:

Manufacturer: Teledyne RDI

Model: Ocean Surveyor

Type: current profiler

Construction:
New items are automatically assigned the state „Construction". These items are listed under „My Devices" (upper navigation banner) for logged-in users with „editor" role."

Public:
Items for which the key metadata fields have been filled and are ready for operation should be assigned the state „Public". The items are listed under „Browse" and „Search" (upper navigation banner) without requiring log-in.

Store:
Items which have been recovered and are available for re-use should be assigned the state „Device Store". These items are listed under „Device Store" (upper navigation banner) for all logged-in users and can also be found under „Search".

Figure 9: Provenance information for instruments registered in Sensor.awi with persistent handles that can be resolved through PANGAEA for dataset-associated instrumentation. Shown here is a) an overview of the metadata components and "status" of instrument panel options, b) exact location of the instrument on the vessel, and c) parameters measured by the Acoustic Doppler Current Profiler installed in the German research Vessel Polarstern (retrieved from sensor.awi.de on 28.01.2019).

The mutual linking of instrument metadata and dataset metadata will be prototyped using datasets submitted from the MOSAiC expedition[34], where the ice-faring research vessel "Polarstern" will be trapped in the Arctic Ice sheet for a year, starting in the fall of 2019. For this expedition, all primary instrumentation will be registered with Sensor.awi and data will be submitted to PANGAEA. This gives us a unique

[34] Mosaic Expedition: https://www.awi.de/en/focus/mosaic-expedition.html

opportunity to test the linkage of data and instruments through PANGAEA and Sensor.awi, and to demonstrate the advantages of linking data and instruments, as an important part of the overall PID Graph.

6.4 Summary table

Definition	For PANGAEA, the core definition of provenance is information that identifies the origin and history of a dataset facilitating full transparency of changes through time.
Purpose	Facilitating full transparency and traceability of the history of curated and published datasets for the data user of the PANGAEA database. Versioning of datasets is an essential part of the provenance, keeping record of changes to a dataset, even past the initial DOI assignment.
Method	If a dataset is published, it cannot be changed anymore. Its integrity can be controlled by using a version identifier which is unique to the content of the dataset and which changes whenever the content is modified. Hence, if the content has changed, an updated DOI assignment is required. In cases where only the format or metadata, but not the content has changed, PANGAEA has several ways to refer to previous versions in different formats.
Standards (current)	PANGAEA metadata is stored in a proprietary (internal format), which is extensible. Its main purpose is to allow converting it to several formats, including DataCite, DublinCore, schema.org or community specific formats like ISO-19139 without information loss. Metadata updates are pushed automatically to DataCite where previous versions of the metadata remain (WP2, T2.2).
Standard integration (future)	PANGAEA metadata will be extended to include provenance information (e.g. identifiers for instruments and platforms). Mapping from the internal PANGAEA schema to DataCite and schema.org metadata needs to be developed.
Implications for the PID Graph	PANGAEA is working on expanding the amount of information provided as metadata. This includes metadata for instrumentation. In a larger perspective, this is also part of a dataset’s provenance, providing information about the procurement of the dataset. The PID Graph can be used to link datasets with specific instrumentation used for their procurement.

7 STFC

7.1 General approach to provenance

Large-scale scientific facilities are expensive. The European Spallation Source under construction in Sweden has a construction budget of €1843 million. At national scale, the Target Station 2 for the ISIS Neutron and Muon Source (an STFC facility in the UK) cost £145 million in 2009. The costs are not only a one-off investment in construction, but ongoing costs of maintenance and upgrading, including the addition of new instruments from time to time. Governments and other funding bodies that pay for these facilities expect some evidence of value for money, and the way this is presented is generally by examining the "impact" of the facilities through the scientific results that are produced through their use. The connection between these results, taking the form of the standard research outputs of publications and datasets, and the "return on investment" is highly complex, and well outside the scope of FREYA; but the point is that it is important to make available reliable information about the outputs and their attribution to particular facilities and instruments.

Such attribution is a type of provenance of the output. It corresponds to the definition given in the introduction to this deliverable: "systematic management of the records of origin of research artefacts". The qualification "systematic" is important: in order to be a reliable basis for whatever impact analysis is to be performed, the recording of attribution must be a regular and sustained part of the procedures of the facility. Suppose that a connection is to be made between the grants of a funding agency and the usage of a facility on those grants: a chain could be established (indeed a small PID graph) between the funding body, the grant, the publications resulting from that grant, the datasets that support those publications, and the instruments on which the data was taken. If the last link is incomplete, then the facility is missing information that supports its mission and justifies its existence.

This variety of provenance is clearly the "provenance of other research artefacts with PIDs" identified in the introduction as one of the flavours of provenance. The question arises, is there also a role for provenance of PIDs themselves and their metadata? Certainly, an aspect of the reliability of attribution is its trustworthiness: how were the links between datasets and facilities asserted and is that trustworthy? The concern is not so much deliberate falsification, which is highly unlikely and easily detected by other means, but assurance of the robustness of the processes in general. This is a secondary concern, however.

7.2 Current provenance activities and implementations

At STFC, the ISIS facility assigns persistent identifiers (specifically, DataCite DOIs) to "investigations" - an investigation being a grouping of associated experiments for a particular research purpose (and therefore leading to multiple datasets). The DOIs resolve to a landing page which is populated with metadata collected from the research proposal managed by a facility-specific proposal system, and then links to the datasets as they are taken on the instruments as part of the investigation. The landing page also includes a reference to the particular instrument used, though this is simply a URL to a local page rather than a PID. It is therefore difficult to systematically make connections between instruments and outputs, though the required information is present. It is probably true to say that the thinking behind these links is not driven by provenance, though the concept is in the background.

No provenance-related standards are currently used in STFC's facilities. In principle the PROV model could be used as a high-level representation. The entities of PROV would be the artefacts and resources arising during the research lifecycle such as datasets and publications. Agents would correspond to particular researchers, instruments, or the facility as a whole, as well as software used for processing and analysis. The activities would refer to gathering of data and processing of it. A complex chain of provenance could be represented in this way. This does not mean that all these aspects have to be modelled in order to construct the PID Graph.

7.3 Plans and considerations for the future

The focus of STFC work on provenance in FREYA is concentrated in another Work Package on new PID types; it started on M16 and will be reported in D3.3 (M27). Research provenance considerations have been taken into account for the pilot application under development that reflects on the PhD research case and is reported in D4.3. By combining metadata from a few diverse repositories in the knowledge graph, it has been possible to augment a path between a facility and a paper resulted from facility research with the (currently free-text) reference about the facility instrument actually used.

A very desirable step towards realising provenance through PID graphs for facilities science would be to have PIDs for instruments on facilities. There are a number of initiatives working towards this goal, including an ORCID User Facilities and Publications Working Group[35], and an RDA Working Group on Persistent Identification of Instruments[36]. Depending on the progress with instrument PIDs implementation, the instrument information that as mentioned above is currently a free-text attribute of certain relations in the knowledge graph, can be further abstracted and represented as a separate Instrument node with a clear identity; this will provide then a richer provenance context for research papers and datasets.

7.4 Summary table

Definition	For STFC (and facilities science in general), provenance relates to resources, that is, to the origin and history of particular entities arising in the research lifecycle. It is worth noting, though, that if provenance is to be used for assessing impact, there should be assurance that the basis of the provenance is trustworthy - so provenance in the sense of “who made this assertion” may also be important.
Purpose	A number of STFC-associated user stories relate to provenance (and these will be generally relevant across facilities science). The unifying theme is one of attribution: associating some outputs with the facility or instrument that gave rise to them. The motivation may be for impact assessment, or for assurance of the origins of the output (e.g. how a published paper made use of the facilities to generate data, and perhaps of software to process it).
Method	This type of provenance may be captured in and emerges from a PID graph, and that is the appropriate level for modelling it. There are questions of appropriate levels of granularity of the entities that the PIDs identify.
Standards (current)	No provenance-specific standards are currently used.
Standard integration (future)	Suitable PIDs for instruments (facilities, beamlines) and software, and the associated metadata, are of course necessary for many use cases where attribution is required.

[35] User Facilities and Publications Working Group: https://orcid.org/content/user-facilities-and-publications-working-group

[36] RDA Persistent Identification of Instruments Working Group: https://www.rd-alliance.org/groups/persistent-identification-instruments-wg

Implications for the PID Graph	A particular PID Graph captures provenance in the sense of attribution and is a basis for queries providing information on the outputs and impact of facilities and beamlines.

8 Discussing provenance approaches

The work on provenance within the disciplinary pilot applications resulted in detailed discussions about definitions, standards, purpose, metadata, etc. and, with that, vastly influenced the definition of the core services in WP2 and potential integrations within the pilot applications. Based on the previous chapters, certain observations should be noted by comparing the approaches described.

First, the discussions and work on provenance revealed commonalities in the understanding and definitions of provenance, i.e. resource provenance vs. metadata provenance, in the different pilot applications of FREYA. It emerged that most pilot applications understand provenance as information about a resource, which is captured in metadata; this refers to documenting the journey of a digital object from inception to processing to publication and the parties involved in this process. Resource provenance is the focus of the majority of the pilot applications with the exception of DANS where provenance primarily refers to the origin of metadata (metadata provenance).

This connects to the reason why the communities study provenance and work on it actively: most of the pilot applications state that provenance information is a means for providing context, for tracking the history of a digital object, as well as who changed it and when. This in turn helps creating trust, rigor and enables reusability. Even more so, provenance information might be relevant or needed for the preservation of audit information (British Library use case).

It could be concluded, that provenance use cases in the FREYA pilot applications focus on addressing a few specific topics: context, trustworthiness, transparency, reusability/ reproducibility, version tracking, assurance of origin, attribution.

All disciplinary partners indicated various standards currently in use that relate to provenance. Due to the diversity of the pilot applications, the differences in standards were expected, as were the disciplinary and "home grown" practices, which were already noted in D4.1. The British Library, CERN, PANGAEA and STFC do not use any "formal", provenance-specific standards. However, that does not exclude provenance information from being exposed to the user through using other standards or custom metadata solutions.

Based on the evaluation of the current state and disciplinary standards, common future plans were brought to the table and were heavily discussed, e.g. PROV and the recent implementation by DataCite (D2.2).

This has led to considerations regarding future integration of provenance standards and provenance-related services. We do see that some partners are moving into the direction of integrating the work done by DataCite in WP2, i.e. incorporating some PROV terminology. It does not seem likely that disciplinary partners will go beyond the integration of the standard service that will be provided by DataCite at least with regard to PROV. DANS is looking into PROV for their JSON for linked data. Some partners are exploring extensions of their work on schema.org, which is not directly connected to provenance, but schema.org can be used to capture provenance information in a less granular way. At EMBL-EBI, discussions are ongoing and more W3C standards are in use and explored; it is not clear yet whether PROV will be integrated.

As for how PIDs fit in the provenance discussion, on the one hand, they (with their metadata) can support providing context, i.e. by being interconnected in the PID Graph; on the other hand, PIDs can capture provenance themselves in their metadata and with that can help enriching the PID Graph content. So, while the work in WP2 is concerned with provenance of DOI metadata, the work in WP4 focuses on provenance for the actual resources or the metadata of the resources the PIDs identify.

The most crucial point in the discussion on provenance is how that relates to PIDs and the FREYA PID Graph. The discussions underlined that provenance is considered a way to enrich the PID Graph. That happens by using and connecting PIDs, where related PID-identified resources (PID Graph concept) or PID metadata itself can provide valuable context to the original object. Connecting to services like DataCite that collect metadata about PIDs, already ensures that important provenance information is captured for DOIs without having to build something on the client’s side.

9 Conclusions

There have been many discussions in the FREYA project about provenance, its meaning for PID core services and the disciplinary pilot applications. This is an important discourse to have across communities to understand the impact of the different types of provenance and their role in serving our communities and EOSC.

Just as Deliverable D4.1 highlighted, these considerations show a great variety of approaches and the key question is whether this is an asset or challenge for the work on the PID Graph, FREYA and EOSC. We reflected on the approaches by each disciplinary partner and concluded that the diversity in the approaches for provenance is an asset for the growing PID Graph and for EOSC, as long as we continue working on exposing information in a machine-readable way that allows interoperability.

The discussions concluded that this needs some attention and work within the communities and by the core service providers (WP2). Core services can provide complementary information so that not every service provider needs to capture every piece of provenance by themselves but can also connect and profit from central services (e.g. by using DataCite DOIs, they can have DOI metadata captured through DataCite's work on provenance). From the research communities themselves we have learnt that resource provenance can help creating trust in certain infrastructures, which could be relevant for the uptake of EOSC.

A condensed version of the discussion has been presented above. The discussion underlined that - despite many solutions on the table - we are still in the beginning of a journey and this deliverable is the first building block in our ongoing work on provenance. This has many different explanations which, again, lie in the different histories of each community, service and organisation. By examining the current state and future, the discussion about provenance showed that is has the potential to enrich the services and PID Graph with valuable information for their user communities and partnering service providers. Moreover, through regular joint meetings on this topic, the discussion helped surfacing needs and requirements that informed the core service development for the PID Graph.