Provenance

Introduction defining context and scope

Provenance, deriving from the French term ‘provenir’ with the meaning ‘to come from’, was originally used to keep track of the chain of ownership of cultural artefacts, such as paintings and sculptures as it determines the value of the artwork. But this concept becomes more and more important also in the data-driven scientific research community. Here it is used synonymously with the word lineage meaning origin or source. The knowledge about provenance of data produced by computer systems could help users to interpret and judge the quality of data a lot better. In the W3C PROV[1] documents provenance is defined as information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

---

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

Change history and amendment procedure

The review of this topic will be organised by in consultation with the following volunteers: . They will partition the exploration and gathering of information and collaborate on the analysis and formulation of the initial report. Record details of the major steps in the change history table below.For further details of the complete procedure see item 4 on the Getting Started page.

Note: Do not record editorial / typographical changes. Only record significant changes of content.

Date	Name	Institution	Nature of the information added / changed

Sources of information used

As this topic is intensively studied both from the research viewpoint and from the viewpoint of those deploying and using provenance in production contexts, there are a large number of relevant papers and reports, cited from the text and further identified in the Reference Section of the deliverable D5.1. The following urls identify other useful sources:

- https://www.w3.org/TR/prov-overview
- https://rd-alliance.org/groups/research-data-provenance.html
- https://rd-alliance.org/sites/default/files/Krohnposter.pdf
- http://de.slideshare.net/drshorthair/om-alignment-with-ssn-prov-oboe-bfo
- http://twiki.ipaw.info/bin/view/Challenge/WebHome
- https://github.com/NCEAS/open-science-codefest/wiki/ProvenanceR
- https://www.dataone.org/webinars/provenance-and-dataone-facilitating-reproducible-science
- http://eprints.soton.ac.uk/271449/1/opm.pdf
- http://d2i.indiana.edu/provenance
- https://kepler-project.org/users/add_on_modules/provenance
- http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
- http://www.mygrid.org.uk/projects/semantic-provenance-project/
- https://www.eudat.eu/semantics
- http://sead-data.net/

Two-to-five year analysis

Already by early 2000, provenance of the scientific results was regarded as important as the result itself. [Moreau 2007] considers that, in order to support reproducibility, workflow management systems are required to track and integrate provenance information as an integral product of the workflow. Consequently [Tan 2007] distinguishes between workflow provenance (or coarse-grained), which refers to the record of the entire history of the derivation of the final output of the workflow and data (or fine-grained) provenance, which gives a detailed account of the derivation of a piece of data that is in the result of a transformation step specified in a database query. [Krohn 2014] calls the latter the database provenance with its sub concepts *why*, *where* and *how provenance*. These describe relationships between data in the source and in the output, for example, by explaining *where* output data came from in the input [Bunemann 2001], showing inputs that explain *why* an output record was produced [Bunemann 2001] or describing in detail how an output recording was produced [Cheney 2009]. [Krohn 2014] adds to this characterisation a third type – provenance of web resources with its sub concept access provenance including both actions of publication and consumption of data. [Hartig 2009] provides a base for research on the provenance of linked data from the Web. [Park 2008] describes republishing as the process of transforming sensor data across the Internet. [Lebo 2014] introduces PROV Pingback which enables parties to discover what happened to objects they created after they have left their domain of influence following the Linked Data principles.

Researchers still face the challenging issue that the provenance of the data products they create is often irretrievable. In many cases the tools for composing lineage metadata are not provided with the software used for scientific data processing. [Bose 2005] sees also the problem that no definitive method, standard or mandate exists for preserving lineage of computational results. While this was true in the early 2000 the provenance community reached a significant milestone in 2013 when the World Wide Web Consortium (W3C) published its PROVenance documents. Although combining PROV with Linked Data offers great potential for discovery, access and use of provenance data, the research community needs practical answers about how to do it. Solutions are necessary to bridge the gap between existing systems built on technologies not well suited to adopting Linked Data design and an interconnected Web of provenance with other systems [Lebo 2014]. [Stehouwer 2014] comes to the same conclusion: there seems to be consensus that it would be very good to move away from manually executed or *ad-hoc*-script-driven computations to automated workflows, but there is still a reluctance to take this step. Traditional approaches of provenance management have focused on only partial sections of data lifecycle and they do not incorporate domain semantics, which is essential to support domain-specific querying and analysis by scientists [Sahoo 2011]. Often analysis has to be performed on scientific information obtained from several sources and generated by computations on distributed resources. This unleashes the need for automated data-driven applications that also can keep track of the provenance of the data and processes with little user interaction and overhead [Altintas 2006]. Comprehensive provenance frameworks as proposed by [Sahoo 2011], [Garijo 2014a], [Myers 2015] or [Filgueira 2015] seem to be the adequate answer to overcome these challenges. These approaches differ from each other and are described below in more detail.

The following section specifies some basic issues related to provenance (see Simmhan 2005): uses, subject, representation, storage, dissemination, tools, collection supported by scientific workflows and by semantic based provenance systems.

Different **uses of provenance** can be envisaged, while currently specific provenance systems typically only support a couple of them [Simmhan 2005]:

Data quality: Lineage can help to estimate data quality and data reliability based on the source data and transformations. It is also used for proof statements on data derivations.

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 its publication, which could expose sensitive and personal information. 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].

Replication recipes: detailed provenance information can allow repetition of data derivation.

Attribution: pedigree of data can give credit and legal attribution to the data producers, enable its citation and determine liability in case of erroneous data. Summaries of such records are useful when funders review the value of continuing support for data services.

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.

The **subject of provenance** information can be of different types as already mentioned above depending on its transparency:

Data-oriented provenance is gathered about the data product and is explicitly available.

Process-oriented (deduced indirectly) provenance focuses on the deriving processes inspecting the input and output data products.

The ***granularity*** at which provenance is detected determines the cost of collecting and storing the related information. The range spans from provenance on attributes and tuples in a database to provenance of collections of files.

Representation of Provenance: different techniques can be used depending on the underlying data processing system.

Annotation: metadata including derivation history of a data product is collected as annotations and descriptions. This information is pre-computed and thus readily usable as metadata.

Inversion: derivations can be inverted automatically to find the source data supplied to them to derive the output data e.g., queries, user-defined functions in databases. This method is more compact.

Provenance related metadata is either directly attached to a data item or its host document or it is available as additional data on the Web [Hartig 2009]. Both types may be represented in RDF using vocabularies or it may be data of another form. The most common *representation languages* used are

- XML
- RDF/OWL using domain ontologies
- CERIF
- dispel4py

Various vocabularies and ontologies exist that allow users to describe provenance information with RDF data.

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 2007). After the Third Provenance Challenge, the Open Provenance Model (OPM) consolidated itself as the *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[1], which are a set of W3C recommendations on how to model and interchange provenance in the Web.

*OPM**[2]***: In OPM (Open Provenance Model) provenance is represented by graphs. It is used to describe workflow executions. The nodes in this graph represent three different types of provenance information: resources created as *artefacts* (immutable pieces of state), steps used as *proces ses* (actions or series of actions performed on artefacts) and the entities that control those processes as *agents*. The edges are directed and have predefined semantics depending on the type of their adjacent nodes: *used* (a process used some artefact), *wasControlledBy* (an agent controlled some process), *wasGeneratedBy* (a process generated an artefact), *wasDerivedFrom* (an artefact was derived from another artefact) and *wasTriggeredBy* (a process was triggered by another process). *Roles* are used to assign the type of activity that artefacts, processes and agents played in their interaction and *accounts* are particular views on the provenance of an artefact. 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.

The *PROV* model is very much influenced by OPM. Here resources are modelled as *entities* (which can be mutable or immutable), the steps used as *activities*, and the individuals responsible for those activities as *agents*. Seven types of relationships are modelled: *used* (an activity used some artefact), *wasAssociatedWith* (an agent participated in some activity), *wasGeneratedBy* (an activity generated an entity), *wasDerivedFrom* (an entity was derived from another entity), *wasAttributedTo* (an entity was attributed to an agent), *actedOnBehalfOf* (an agent acted on behalf of another agent) and *wasInformedBy* (an activity used an entity produced by another activity). Roles are kept to describe the type of relationship and the means to qualify each of the relationships using an n-ary pattern are provided. OPM introduces the concepts *plan* associated with a certain activity and PROV statements grouped in *bundles* defined as entities.

Figure 10: The communalities between PROV (left) and OPM (right) [Garijo 2014a].

The PROV family of documents provides among others an ontology (PROV-O), the data model (PROV-DM) and an XML schema (PROV-XML).

Provenir [Sahoo 2011]: is a domain-upper ontology provenance ontology used in translational research. 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 along with ten fundamental relationships. The three top-level classes are data, process and agent, where data is specialised in the classes data_collection and parameter (spatial, temporal and thematic). Provenir is used in the semantic provenance framework (SPF) as explained below.

P-PLAN [Garijo 2014a]: in order to be able to represent workflow templates and workflow instances [Garijo 2014a] extended PROV. 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 *variabl e* with the properties: type, restrictions and metadata.

OPMW [Garijo 2014a]: is designed to represent scientific workflows at a fine granularity. OPMW extends 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. Additionally, it supports representation of attribution metadata about a workflow. OPMW is used as provenance representation model in the WEST workflow ecosystem.

*O&M alignments with PROV**[3]***: To be compliant with the OGC standard ISO 19156 (Geographic Information – Observation and Measurement) Simon Cox (2015) made efforts to align O&M with PROV. In O&M an observation is an action whose result is an estimate of the value of some property of the feature-of-interest, obtained using a specified procedure.

Provenance storage: in case the data is fine grained, provenance information can become larger than the data it describes. This determines its scalability. This is particularly true when annotations are added manually instead of automatically collecting them.

Provenance dissemination: In order to use provenance, a system should allow rich and diverse means to access it. These can include provenance mining, visualisation and browsing. If provenance is stored in RDF/OWL it is possible to query using SPARQL. Many tools have been developed for PROV for this purpose. A visualisation tool like PROV-O-viz produces derivation graphs that users can browse and inspect [Garijo 2014a].

Provenance collection: might be performed by stand-alone tools such as ProvenanceR[4], which enables provenance capture in R but these are of more useful when **embedded in a workflow system.**

Provenance collection supported by scientific workflow systems: data analysis can be facilitated by scientific workflow systems that have the ability to make provenance collection a part of the workflow. Here the provenance should include information about the context in which the workflow was used, execution that processed the data and the evolution of the workflow design. Among the most popular of these are Taverna, Kepler and Pegasus. Here only a few a described in some detail – see also descriptions in Section 3.5.

Kepler: is a cross-project collaboration to develop a scientific workflow system for multiple disciplines that provides a workflow environment in which scientists can design and execute workflows. Kepler uses Ptolemy II software, a Java-based system and a set of APIs. The focus is to build models based on the composition of existing components, called 'actors', and observe the behaviour of these simulation models when executed using different computational semantics called 'directors'. Formerly a Provenance Recorder had been implemented to be configured as a 'director' with a standard configuration menu and becoming part of the workflow definition [Altintas 2006]. Today the Kepler Provenance enriches the capabilities of the workflow as add-on module suite. Provenance is toggled on and off in the Kepler toolbar. When on and when running a workflow with a supported director (SDF, DDF, or PN), execution details are recorded into a database in the KeplerData/modules /provenance directory. This powerful feature is leveraged by modules such as Reporting and the Workflow Run Manager, which provides a GUI to manage and share your past workflow runs and results[5].

The *dispel4py* data-streaming system [Filgueira 2015], [Spinuso 2016]: is a versatile data-intensive kit presented as a standard Python library. It describes abstract workflows for stream-based applications, which are later translated and enacted in distributed platforms. It allows users to define abstract, machine-agnostic, fine-grained data-intensive workflows. Scientists can easily express their requirements in abstractions closer to their needs without demanding knowledge of the hardware or middleware context in which they will be executed. A processing element (PE) is a computational activity. It encapsulates an algorithm or a service, and is instantiated as node in a workflow graph. Users only have to use available PEs from the dispel4py libraries and registry, and connect them as they need in graphs which leads to extensive re-usability. The provenance management system of dispel4py consists of a comprehensive system which includes extensible mechanisms for provenance production, a web API and a visualisation tool. The API is capable of exporting the trace of a run in the W3C-PROV JSON representation to facilitate interoperability with third party tools.

Provenance collection supported by semantic-based provenance systems:

Taverna: is an open source and domain-independent Workflow Management System comprising a suite of tools to design and execute scientific workflows. It has been created by the myGrid team and is funded by FP7 projects BioVeL, SCAPE and Wf4Ever. It is written in Java and includes the Taverna Engine (used for enacting workflows) that powers both Taverna Workbench (the client application) and Taverna Server (executing remote workflows). Taverna automates experimental methods through the use of a number of different services from a diverse set of domains. It enables a scientist who has a limited background in computing, limited technical resources and support, to construct highly complex analyses over data and computational resources. Workflow sharing is arranged via myExperiment. Taverna can capture provenance of workflow runs, including individual processor iterations and their inputs and outputs. This provenance is kept in an internal database which is then used to populate the history results in the results perspective in the Taverna Workbench. The provenance trace can be used by the Taverna-PROV plugin to export the workflow run, including the output and intermediate values, and the provenance trace as a PROV-O RDF graph which can be queried using SPARQL and processed with other PROV tools, such as the PROV Toolbox. Within Taverna, a workflow can be annotated to give attribution to the Authors of a workflow (or nested workflow)[6]. Although Taverna is not semantic based it supports the semantic description of workflows.

The semantic provenance framework (SPF) (Sahoo 2011): provides a unified framework to effectively manage provenance of translational research data during pre and post-publication phases. It is underpinned by an upper-level provenance ontology (Provenir) that is extended to create domain-specific provenance ontologies to facilitate provenance interoperability, seamless dissemination of provenance, automated querying with SPARQL and analysis. To collect provenance information at a first stage existing data stored in RDB was converted to RDF with help of D2RQ using the domains-specific Parasite Experiment ontology (PEO). On a second stage an ontology-driven web form generation tool called Ontology-based Annotation Tool (OntoANT) was developed to dynamically generate web forms for use in research projects to capture provenance information consistent with PEO in RDF. The SPF stores both the dataset and provenance information together in a single RDF graph. This allows for application-driven distinction between provenance metadata and data, and additionally facilitates that updates of data are seamlessly applied to the associated provenance.

The WEST workflow ecosystem [Garijo 2014a]: integrates different workflow tools with diverse functions (workflow design, validation, execution, visualisation, browsing and mining) created by a variety of research groups. Workflow representation standards and semantic technologies are used to enable each tool to import workflow templates and executions in the format they need. WEST uses and extends the Open Provenance Model and the W3C PROV standard by P-Plan which is able to represent plans. The extension is considered necessary because the OPM and PROV models are not able to represent workflow templates and workflow instances. The OPMW vocabulary is designed to represent scientific workflows at a fine granularity built upon P-Plan, OPM and PROV, and allowing the linking between a workflow template, a workflow instance created from it, and a workflow execution that resulted from an instance. [Garijo 2014a] demonstrate the efficiency of such an approach by the usage of different tools such as WINGS for generating workflows, workflow execution engines such as Pegasus, the FragFlow system for workflow mining, Prov-o-viz for visualising provenance structures, WExp for exploring different workflow templates, the Organic Data Science Wiki, an extension of semantic wikis for workflow documentation and Virtuoso as workflow storage and sharing repository.

Life Science Grid (LSG) (Cao 2009): is a cyber-infrastructure framework supporting interactive data exploration and automated data analysis tools. It uses the Karma provenance framework[7] developed at Indiana University to capture raw provenance events and to format them according to the Open Provenance Model specification. Additionally, it integrates automated semantic enrichment of the collected provenance metadata using the Semantic-Open Grid Service Architecture (S-OGSA) semantic annotation framework developed at University of Manchester.

*The Sustainable Environmental Actionable Data (SEAD)**[8]**:* provides data curation and preservation services to deploy those services for beneficial use to active research groups. It intends to support the 'long-tail' of smaller projects in sustainability science. Assuming that metadata could be used to help organise and filter data during research, the SEAD approach allows data and metadata to be added incrementally, and the generation of citable persistent identifiers for data. It comprises three primary interacting components: Project Spaces, Virtual Archive and Researcher Network. The Project Space is a secure, self-managed storage with tools that allow research groups to assemble, semantically annotate and work with data resources. The web application leverages the Tupelo semantic content middleware developed at NCSA, which provides a blob plus RDF metadata abstraction over an underlying file system and RDF store. The web application itself is an extension to the Java-based Medici semantic content management web application. SEAD has also added a set of restful web services that can be used within the R analysis application to read and write data with desired provenance and metadata. A SPARQL-query service is also implemented. The Virtual Archive is a service that manages publication of data collections from Project Spaces to a range of long-term repositories. It is a federated layer over multiple repositories that manages an overall packaging and publication workflow and provides a global search capability across data published via SEAD. It leverages the Komadu provenance service[9] which is a stand-alone provenance collection tool that can be added to an existing cyberinfrastructure for the purpose of collecting and visualising provenance data. It supports the W3C PROV specification. Komadu is the successor of the Karma provenance tool which is based on OPM.

Another semantic tool which can be adopted for provenance information collection is B2NOTE[10]: The EUDAT project developed a first prototype version using python and common semantic python libraries like RDFlib and SPARQLWrapper. This webservice allows annotation of imported text/documents with terms coming from Bioportal, EnvThes and GEMET from EIONET. This prototype is currently being tested and extended using the Django RESTful framework to be further integrated with the LTER/LifeWatch portal.

---

[1] https://www.w3.org/TR/prov-overview/
[2] http://eprints.soton.ac.uk/271449/1/opm.pdf
[3] http://de.slideshare.net/drshorthair/om-alignment-with-ssn-prov-oboe-bfo
[4] https://github.com/NCEAS/open-science-codefest/wiki/ProvenanceR
[5] https://kepler-project.org/users/add_on_modules/provenance
[6] http://www.taverna.org.uk/documentation/taverna-2-x/provenance/
[7] http://d2i.indiana.edu/provenance_karma
[8] http://sead-data.net/
[9] http://d2i.indiana.edu/provenance_komadu
[10] https://www.eudat.eu/semantics

Sketch of a longer-term horizon

- In order for data-driven research to be reproducible it is an essential requirement to define unambiguously all data inputs, analysis steps and data products, as well as software and algorithms used with *persistent identifiers*. This will allow for connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows.
- Future provenance management developments will have to implement *interoperability* functions of workflows. The need for global inter-disciplinary collaborations will continue to grow with demands for scientific data to be shared, processed and managed on different *distrib uted computational infrastructures*.
- Provenance management should embrace the *whole life cycle* of data and incorporate *domain semantics* by encouraging and building on controlled vocabularies formalised as ontologies – see Section 3.9, which is essential to support domain-specific querying and analysis by scientists. The approach used for provenance representation has a significant impact on the storage, dissemination, and querying phases of the provenance life cycle [Sahoo 2011].

- Provenance analytics and visualisation techniques will receive more attention in future applied research [Spinuso 2016]; so far it has been largely unexplored. By analysing and creating insightful visualisations of provenance data, scientists can debug their tasks and obtain a better understanding of their results. [Davidson 2008], [Cao 2009].

Relationships with requirements and use cases

Requirements:

There is a big interest among the RIs to get clear recommendations from ENVRIplus about the information range provenance should provide. This includes drawing an explicit line between metadata describing the 'dataset' and provenance information. Also it should be defined clearly whether usage tracking should be part of provenance.

It is very important to provide support for automated tracking solutions and provenance management APIs to be applied in the specific e-science environments. Although there are some thesauri already in use there is a demand for getting a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes.

Work Packages:

There is a strong relationship between WP 6 and the WP 8 task 3 *Provenance* as there must be a direct link between the data and its lineage that can be followed by the interested user. The recommendations provided for data identification and citation should be used in provenance service solutions. Provenance tracking is also an important feature for the tasks 7.1 processing and 7.2 optimisation. The connections with the tasks 8.1 curation and 8.2 cataloguing are evident as well as all of these recommendations must be built upon the same data model, semantically and technically speaking, as defined in the task 5.3 semantic linking framework and integrated in the task 5.4 interoperation based architecture design.

Relationships with use cases as foreseen in WP9:

- **IC_1, Dynamic data citation**: Connections to cataloguing and maintenance of provenance records, supporting automated metadata extraction and production for machine-actionable workflows
- **IC_2, Provenance:** aims amongst others at defining a minimum information set that has to be tracked, finding a conceptual model for provenance which conforms to the needed information, maps existing models to the common model, and finds a repository to store the provenance information.
- **IC_06, Identification/citation in conjunction with provenance**: is aimed at identifying good practices for using PIDs for recording provenance throughout the data object lifecycle, including workflows and processing.
- **IC_8, Cataloguing, curation and provenance:** is the implementation case for catalogues fulfilling curation and provenance requirements.
- **IC_9, Provenance – use of DOI for tracing data re-use:** provenance capture techniques will be used as background for this use case.
- **IC_11, semantic linking framework:** interoperability and semantic linking across catalogues (e.g., datasets with observation systems and persons) upon a common data and metadata model will be provided by this use case.

Summary of analysis highlighting implications and issues

- Commonality of metadata elements across curation, provenance, cataloguing (and more) thus a common metadata and provenance scheme based on widely adopted international standards should be used.
- Link to existing vocabularies and ontologies to enable domain semantic provenance representation thus a strong collaboration with the semantic working group.
- Having better visualisation tools at hand for provenance dependencies will increasingly help to reduce the RIs reluctance to adopt workflow solutions with provenance functionalities – thus it is important to follow related developments and to try to implement the most relevant one(s) in the provenance service.
- ENVRIplus should consider collaborating with EUDAT on the development of provenance tools as foreseen in WP 8 and influence the General Execution Framework (GEF) so that it supports the provenance-collection functionality.
- ENVRIplus should follow the RDA provenance working groups and participate.
- Provenance in ENVRIplus is a task which is due in a later stage of the project. Thus it is a must to follow in the meantime tools and services now under development that will allow seamless linking of data, articles, people supporting streamlining of the entire data management cycle, virtually instantaneous extraction of metadata and provenance information, and facilitating data mining and other machine-actionable workflows.

Further discussion of the provenance technologies can be found in Section 4.2.9. This takes a longer term perspective and considers relations with strategic issues and other technology topics.

Bibliography and references to sources

[Altintas 2006] Altintas, I., Barney, O. Jaeger-Frank, E.: *Provenance Collection Support in the Kepler Scientific Workflow System*. L. Moreau and I. Foster (Eds.): IPAW 2006, LNCS 4145, pp. 118-132, 2006.

[Bose 2005] Bose, R., Frew, J. *Lineage Retrieval for Scientific Data Processing: A Survey*. ACM Computer Surveys, Vol. 37, No. 1, 2005.

[Bier 2013] C. Bier: *How usage control and provenance tracking get together – a data protection perspective.* Security and Privacy Workshops (SPW), 2013 IEEE, San Francisco, CA, 2013, pp. 13-17.

[Buneman 2000] P. Buneman, Khanna, S., Tan, W.-C.: Data Provenance: Some Basic Issues. Lecture Notes in Computer Science, Volume 1974, Foundations of Software Technology and Theoretical Computer Science, (FST TCS 2000), pages 87-93.

[Buneman 2001] P. Buneman, S. Khanna, and T. Wang-Chiew, *Why and Where: A Characterization of Data Provenance*, in Database Theory — ICDT 2001, vol. 1973, J. Bussche and V. Vianu, Eds. Springer Berlin Heidelberg, 2001, pp 316-330.

[Cheney 2007] Cheney L., Chiticariu L., Tan W-C. *Provenance in Databases: Why, How and Where*. Foundations and Trends in Databases, Vol. 1, No. 4 (2007) 379-474.

[Cao 2009] Cao, B. et al: *Semantically Annotated Provenance in the Life Science Grid.* SWPM'09 Proceedings of the First International Conference on Semantic Web in Provenance Management. Vol. 526, pp 17-22.

[Davidson 2008] Davidson, S.B., Freire, J.: *Provenance and Scientific Workflows: Challenges and Opportunities*. SIGMOD'08, Vancouver, Canada.

[Duerr 2011] R.E. Duerr , R.R. Downs, C. Tilmes, B. Barkstrom, W.C. Lenhardt, J. Glassy, L.E. Bermudez and P. Slaughter, *On the utility of identification schemes for digital earth science data: an assessment and recommendations*. Earth Science Informatics, vol 4, 2011, 139-160. Available at http://link.springer.com/content/pdf/10.1007%2Fs12145-011-0083-6.pdf

[Filgueira 2015] R. Filgueira, A. Krause, M. Atkinson, I. Klampanos, A. Spinuso and S. Sanchez-Exposito, *dispel4py: An Agile Framework for Data-Intensive eScience*, e-Science (e-Science), 2015 IEEE 11th International Conference on, Munich, 2015, pp. 454-464.

[Garijo 2014] Garijo, D., Gil, Y., Corcho O.: *Towards Workflow Ecosystems through semantic and standard representations*. Proceedings of the Ninth Workshop on Workflows in Support of Large-Scale Science (WORKS), held in conjunction with SC 2104, New Orleans, LA, November 16, 2014.

[Hartig 2009] Hartig, O., *Provenance information in the web of data*, in Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009), 2009.

[Lebo 2014] Lebo, T., West, P. , McGuiness, D.L.: *Walking into the Future with PROV Pingback: An Application to OPeNDAP using Prizms,* Prove nance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop, IPAW 2014, Cologne, Germany, June 9-13, 2014.

[Lim 2010] C. Lim, S. Lu, A. Chebotko and F. Fotouhi, *Prospective and Retrospective Provenance Collection in Scientific Workflow Environments*, Services Computing (SCC), 2010 IEEE International Conference on, Miami, FL, 2010, pp. 449-456.

[Mureau 2008] Mureau, L. et al: *Special Issue: The first provenance challenge*. Concurrency and computation: practice and experience. 2008: 20, 409-418.

[Myers 2015] Myers, J. et al: *Towards Sustainable Curation and Preservation: The SEAD Project's Data Services Approach*, 2015 IEEE 11th International Conference on eScience, 526-535.

[Park 2008] Park, U., Heidemann, J.: *Provenance in Sensornet Republishing*. J. Freire, D. Koop and L. Moreau (Eds.): IPAW 2008, LNCS 5272, pp. 280-292, 2008.

[Sahoo 2011] Sahoo S.S. et al: *A unified framework for managing provenance information in translational research*. Bioinformatics, 2001, 12: 461.

[Simmhan 2005] Simmhan, J. L., Plale, B., Gannon, D.: *A survey of Data Provenance in e-Science*, SIGMOD Record, Vol. 34, No. 3, Sept. 2005.

[Stehouwer 2014] Stehouwer, H. Wittenburg P.: Second year report on RDA Europe analysis programme, deliverable D2.5.

[Tan 2007] Tan, WC: *Provenance in Databases: Past, Current, and Future*, IEEE Data Eng. Bull.