Computational Provenance in DataONE:

Implications for Cultural Heritage Institutions

Robert J. Sandusky
University Library
University of Illinois at Chicago
Chicago, USA
sandusky@uic.edu

***Abstract*—Provenance data is a type of metadata that computer scientists argue can support trustworthy and reliable replication of scientific results. From its origins in scientific workflow systems and database theory, and with concurrent interest from the ecological informatics community, a standard data model (PROV) and extensions for DataONE (ProvONE) have led to initial implementations in several tools commonly used by scientists (R, MATLAB) and in a global federation of scientific data repositories (DataONE). DataONE's support for ingest, storage, indexing and retrieval of provenance data is presented. Implications for libraries and other cultural heritage organizations are identified. A research agenda for determining the applicability of the PROV model to cultural heritage institutions and their digital asset management systems is presented.**

Provenance; libraries; archives; museums; computational provenance.

I. INTRODUCTION

"Data, the units of information observed, collected, or created during the course of research, is not limited to scientific data, but includes social science statistical and ethnographic data, humanities texts, or any other data used or produced in the course of academic research, whether it takes the form of text, numbers, image, audio, video, models, analytic code, or some yet-to-be-identified data type." [1]

Research data is an integral part of our scientific and cultural history, on a par with scholarly monographs, journal articles, archival collections, poems, and works in the visual and performing arts. In the context of science - including social science - research data forms the evidentiary base of scholarly communication. The ability to trace, in a trustworthy and reliable manner, the conclusions published in a scientific paper back through to the analytic procedures, datasets and data collection procedures has become a fundamental requirement for ensuring public support for science. To cite but two examples where the scientific enterprise has been attacked in order to promulgate certain political or religious agendas, evidence and scholarship overwhelmingly support theories such as anthropogenic climate change and evolution. Despite scientific consensus on these questions, large and sometimes powerful segments of society have made widely-publicized assertions that there is yet no scientific consensus on these topics. An overwhelming 97.4% of climatologists who are active publishers on climate change agreed with the statement Do you think human activity is a significant contributing factor in changing mean global temperatures? while a 2008 public opinion poll in the US showed that about 58% of the general public shares that opinion [2].

How are repository-based research data services providing support for trust, reliability and evidentiary linkage in order to facilitate the replication of scientific results? Provenance data is a type of metadata that computer scientists and ecological informaticists argue can support these needs. Academic and research libraries, information schools, and library and information science (LIS) researchers have been deeply engaged in building services and infrastructure to support research data services for more than ten years [3,4,5,6]. Until now, however, working implementations of provenance for research data and data stewardship have been rare.

This paper describes work done in the context of DataONE, by members of both the computer science and ecological informatics communities, to address the need to provide reliable and trustworthy access to data and digital representations of artifacts, agents, and activities that bridge datasets and scientific publications. Members of these communities also collaborated on the development of World Wide Web Consortium recommendations referred to as PROV. "PROV defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world." [7] This paper explores the implications for organizations operating research data repositories and suggests potential for adoption of PROV in libraries, archives and museums. A research agenda for determining the applicability of the PROV model to cultural heritage institutions and their digital asset management systems is presented.

II. PROVENANCE IN CONTEXT

"Provenance helps determine an object's value, accuracy, and authorship." [8]

The Oxford English Dictionary online defines provenance as "the fact of coming from some particular source or quarter; origin, derivation." [9, meaning 2] The term has other, more specialized meanings in other contexts such as art and cultural heritage, archival science and computer science. In the context of art museums and the practice of collecting cultural heritage artifacts, including paintings and sculptures, provenance is defined as "the

history of the ownership of a work of art or an antique, used as a guide to authenticity or quality; a documented record of this." [9 meaning 3] The trustworthiness of the documented record of an object's ownership is key in the museum context.

Archival science has a different definition, where the word provenance is usually encapsulated in the phrase "the principle of provenance". "The principle of provenance has two components: records of the same provenance should not be mixed with those of a different provenance, and the archivist should maintain the original order in which the records were created…." [10] Here, the notion of "a provenance" refers to an "individual, family or organization". [11] The issue of the trustworthiness of records is still paramount, as "...archivists are concerned that records are trustworthy, that they are what they purport to be." [12] The Society of American Archivists uses the terms "chain of custody" and "custodial history" when referring to "the succession of offices, families, or persons who held materials from the moment they were created". [13]

The notion of computational provenance has emerged from the computer science community during the past fifteen years. Computational provenance researchers seek to develop systematic, computationally-based processes and standards for capturing, and making available for use, information about who created an object, when it was created or modified and the process or procedure that modified the object [8]. Buneman, Khanna, and Wang-Chiew [14] define data provenance as: "...where a piece of data came from and the process by which it arrived in the database…." and suggest the words "lineage" and "pedigree" as synonyms for the term "data provenance".

The development of computational provenance is motivated by the following objectives [8]:

- Produce detailed and trustworthy record keeping in data-driven and big data science
- Automate provenance collection to reduce researcher effort and the probability of introducing errors into the record-keeping process
- Automate analysis tasks
- Help users of existing data interpret and understand reported results
- Improve support for reproducibility in science by sharing details of data manipulation and analysis

Imagine the reader of a journal article who wishes to examine the data originally collected and fully understand the tools and techniques used to process and analyze the data and ultimately produce the paper. Perhaps the reader wishes to apply the same methods to similar data collected in another geographic location. The reader might ask provenance related questions such as:

- Who collected this data?
- When was the data collected?
- Where was the data collected?
- How was the data collected?
- At what point in the research lifecycle was the instance of the data presented in the paper instantiated?
- What procedures were used to clean, normalize, or reduce the data?
- Has this dataset been altered since it was deposited, or since the publication of the paper? If so, by whom, when, and why?
- Are there other data to which this data is related? Is the data reported in this paper a subset of a larger dataset?

Approaches to representing provenance include manually adding logging code to scripts in order to record session data, including parameters and input. However, the data created by this approach can't easily be queried and management of the log files - storing, organizing, and preserving - must be done manually. In addition, if multiple scripts are used, or multiple versions of a script are used over time, log files may not be consistently formatted, complicating their later use, and it may be difficult to track the order in which separate scripts were run [15]. Scientific workflow systems, including Kepler, Taverna, and VisTrails, have been developed to "...provide intuitive visual programming interfaces that are easier to use for people who don't have substantial programming expertise". [15] Workflows are represented by data structures, typically formal graphs [16], which facilitate query and exploration of provenance data.

III. PROVENANCE IN DataONE

A. Design

From the time of its initial design in 2007, DataONE's architecture included support for the storage, preservation and discovery of provenance data in order to provide detailed capture of data lineage [17].

DataONE's Cyberinfrastructure Working Group, the body responsible for the design and implementation of DataONE's technical infrastructure, initially discussed requirements and designs for provenance. Supporting the storage, discovery and preservation of provenance traces generated by scientific workflow systems such as Kepler was always fundamental to DataONE's work on provenance, but issues of the immutability of content and the use of persistent identifiers in DataONE were discussed frequently [18]. An analysis of DataONE's version 1 metadata schema was conducted [19], comparing the draft DataONE system metadata schema with PREMIS 2.0 [20]. DataONE formed a Provenance and Semantics Working Group to develop "...an open and extensible provenance management architecture for scientific data processing systems (e.g., workflows and scripting languages such as R)" [21]. Members of DataONE's Scientific Workflows and Provenance Working Group also participated in the development of the World Wide Web consortium's PROV specification from 2010 through the publication of the specification in 2013 [22]. While much discussion and planning was done early in the project, implementation of provenance was planned for a phase following the initial production implementation of DataONE in July 2012.

The PROV model "…is not tailored to ... any specific scientific application. Instead, it is meant to be generic and accommodate the provenance of data that is generated from a

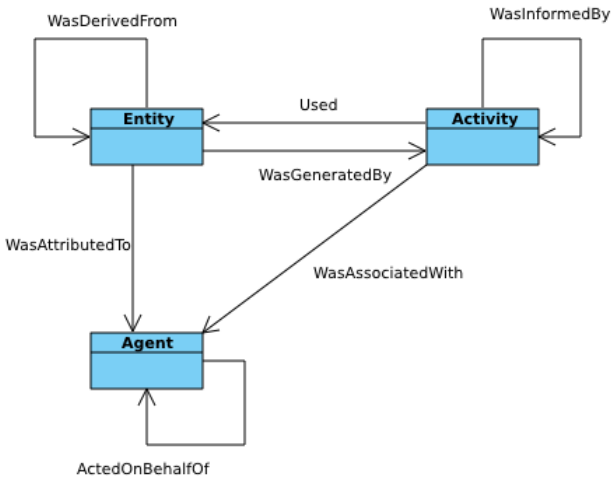

Figure 1. Core types as defined in PROV-DM (n.d.)

variety of diverse data sources, including human information processing" [22]. The model, as documented in the World Wide Web Consortium's documents, is simple. The assertions defined in PROV are designed to support the making of assertions about how objects were "created or delivered"; thus they have names such as wasAttributedTo, wasDerivedFrom or wasGeneratedBy [23]. See Fig. 1 for a diagram of the core types.

The PROV model has been extended by DataONE to support storing and sharing of provenance data created by "scientific workflow-based computational experiments" [24]. The goals of this extension include [24]:

- Support interchange of provenance information generated by different scientific workflow systems which natively use proprietary execution trace formats
- Capture the high-level workflow itself, typically represented as a set of abstract steps or a "recipe" (prospective provenance)
- Capture the inputs, outputs, software agents and intermediate data artifacts involved in the execution of a scientific workflow (retrospective provenance)
- Capture the evolution of a scientific workflow over time (process provenance)
- Address "the most relevant aspects of how the data both used and produced by a computational process…. [T]he inputs and outputs of the various tasks…" (data structure)
- Provide extension points "to accommodate the specificities of particular scientific workflow systems"

Prospective provenance, the representation of the scientific workflow at the level of a set of steps or a "recipe", is expressed in ProvONE by the classes Program, Port, Channel, Controller and Workflow (see Fig. 2, blue/green boxes in the lower left).

Retrospective provenance, the representation of execution traces associated with the use of a particular workflow, is expressed by the classes Execution, Association, Usage, Generation and User (see Fig. 2, golden boxes in the upper left).

Data structures resulting from the execution of a workflow are expressed by the classes Entity, Collection, Data, Visualization and Document (see Fig. 2, purple boxes in the upper right).

Workflow evolution is represented by the wasDerivedFrom association provided with the prospective provenance classes (see arrow in Fig. 2, lower left).

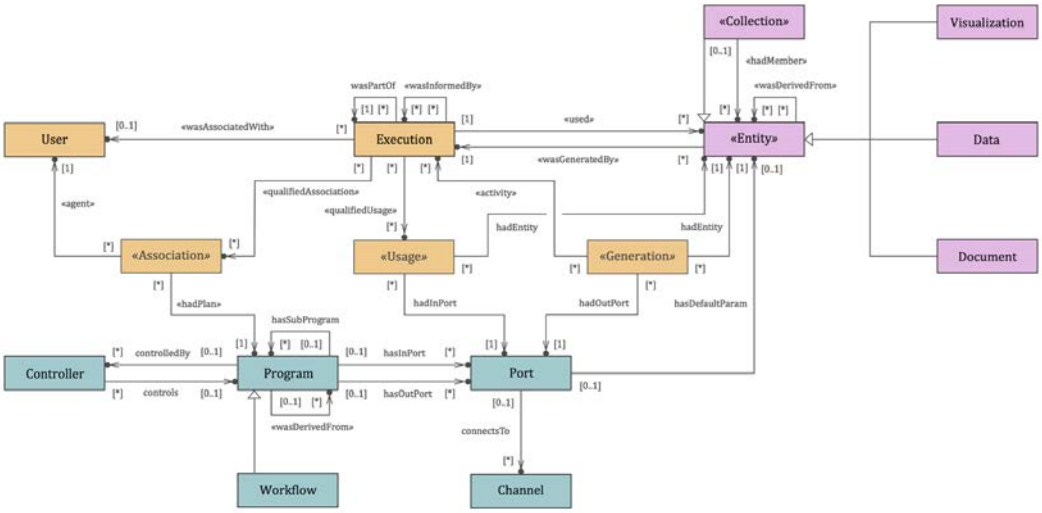

Figure 2. The ProvONE conceptual model extends PROV-DM.

B. *Implementation*

DataONE consists of three kinds of cyberinfrastructure which together facilitate discovery, access, and preservation for scientific data in the earth and environmental sciences. *Member nodes* are existing repositories of scientific data that implement services conforming to the published DataONE API. There are more than 25 member nodes located across the globe (see https://www.dataone.org/current-member-nodes for the current list of DataONE member nodes). Three coordinating nodes located in the US catalog the federated content harvested from the member nodes, provide support for discovery and access, and manage replication of content among geographically dispersed member nodes. The investigator toolkit is a set of new or enhanced software tools that are able to directly contact DataONE and utilize services for authorization, data access, retrieval and deposit [17].

R and MATLAB are two tools in the investigator toolkit to which DataONE has contributed extensions that support direct interactions (e.g., dataset search, dataset ingest, dataset deposit to DataONE) between the tool and the DataONE federated repositories. Both R and MATLAB now have code extensions for provenance [28, 29].

ProvONE provides the conceptual model for representing scientific workflows (prospective provenance) and their evolution, execution of workflows (retrospective provenance) and the data structures that are generated by workflows (documents, datasets, visualizations and collections of these entities). The implementation of the model in DataONE was put into production in the summer of 2015, but its use by scientists is not yet widespread.

DataONE supports search of provenance data by harvesting provenance data from member nodes, representing provenance data internally using OAI-ORE maps [25] and indexing provenance data using Solr. DataONE's search interface provides user interface elements to support exploration and interpretation of the provenance information. At this time, there are few objects with associated provenance deposited into DataONE member nodes, so the basic user interface elements now supported will be described using a demonstration example available online at https://search.test.dataone.org/#view/urn:uuid:bf71c38b-22b2-469e-8983-734ec0ab19cb.

The example [26] is based upon a hydrocarbon database created after the Exxon Valdez oil spill in the Gulf of Alaska in 1989 [27]. This dataset is a data package consisting of eleven elements: one Ecological Markup Data (EML) metadata document, six data files in either .csv or .zip format, two R scripts and two images in .png format (Fig. 3). (There are a number of interesting details in the abstract provided for the dataset [26] including migration of the data from RBase to a proprietary format to Microsoft Access and finally to derived .csv files of the key data tables. The data were also migrated across multiple generations of the Windows operating system.)

Selecting the "More Info" link to the right of "Total_Aromatic_Alkanes_PWS.csv" in the list of files moves the browser window to the metadata for this data file. The circles containing open and closing angle brackets and a slash (</>) are icons representing code or scripts. Selecting either of these icons provides additional detail about the source script that generated the Total_Aromatic_Alkanes_PWS.csv data file and the script that generated the derivations - two image files - based upon this data. The pop-up windows include hyperlinks to support navigation to representations of objects - whether scripts, data files or derivatives - linked by the provenance data. The data file itself can be obtained by selecting the "Download" button (Fig. 4). Selecting the "View >>" icon in the "Derived program" pop-up (Fig. 4) shifts the browser window to the metadata for the "Locations map R script" (Fig. 5).

Fig. 5 shows one data file as input to this R script (Total_Aromatic_Alkanes_PWS.csv) and two images (maps of the Gulf of Alaska) as output. Note that the relationships

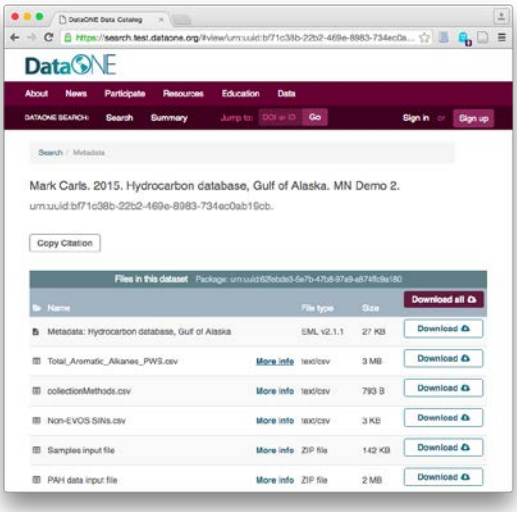

Figure 3. Dataset representation in DataONE search.

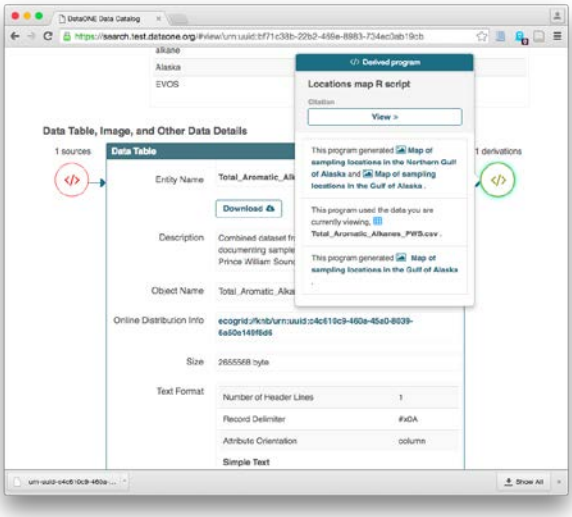

Figure 4. Data file representation including provenance icons for code or scripts.

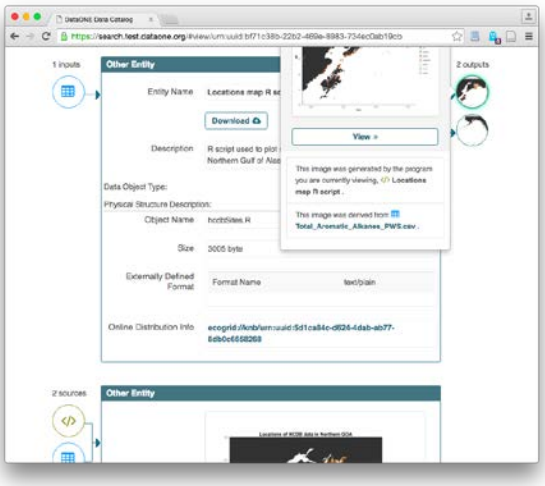

Figure 5. R script representation including provenance icons for one data file as input and two images as outputs; bottom of figure shows the representation for one image, which has one script and one data file as sources, but no outputs.

defined in the ProvONE model are expressed in natural language in the pop-up for the images, signaling to the user that the image was generated by a specific R script and that the image was derived from a specific data file.

IV. IMPLICATIONS FOR CULTURAL HERITAGE INSTITUTIONS

Libraries and research institutions are still working to develop mature research data services, including the infrastructure for repository services [4,5,6]. The development of additional provenance-aware tools and services for scientists, such as DataONE's extensions to R and MATLAB, increases the probability that librarians, publishers and repository managers will begin to field questions about provenance and provenance data or perhaps see deposits of provenance data in repositories. Publishers and repository managers may receive supplemental files that include provenance data, such as instrumented R scripts or Java or Python code, associated log files, or representations of workflows from systems like Kepler or VisTrails. Provenance data is a new form of metadata that is intended to provide assurance that the results of data-driven science, which rely upon the computational generation, transformation, analysis and visualization of data, are reliable and trustworthy.

Deposit of provenance data, whether prospective representations of workflows (the "recipe") or retrospective traces of the execution of a workflow used to generate the figures in a published paper, is not sufficient to make provenance accessible or usable in the future. The implementation of provenance in DataONE, based upon the ProvONE extensions to the World Wide Web's PROV model, has been designed to support the interoperability of provenance data generated by both a range of workflow systems and arbitrary scripts. The DataONE implementation depends upon the use of unique persistent identifiers for objects. ARKs, DOIs or similar persistent identifiers can be used to represent individual data files, sets of data files, and metadata files (sometimes referred to as data packages), scripts (written in R, Java, Python or other languages), or derivations (documents, visualizations, data files). ORCIDs can be used similarly to unambiguously represent human agents responsible creating the data files, scripts, and visualizations included in a workflow. OAI Object Reuse and Exchange (OAI-ORE) provides a way to represent complex provenance networks flexibly in a manner that enables computational approaches to indexing, discovery, and user interface design [25].

Libraries and other cultural heritage institutions, such as museums and archives (LAMs), face challenges similar to supporting reproducibility in data driven science. LAMs are building collections of digitized and born-digital materials, which are generally expected to have a useful life spanning many decades. These institutions have the need to provide trustworthy and reliable digital objects across time and, for preservation purposes, across multiple, geographically dispersed repositories that may be managed by multiple organizations running on a variety of hardware and software. Objects with a multi-decadal lifespan are likely to also be migrated to different media several times, and to be managed by a succession of digital asset management systems during that time.

There are questions to be addressed in both the cases of provenance for scientific data and provenance for LAMs. The most fundamental questions are What is the utility of provenance data? To whom is it useful, how is it useful, and in which contexts? The DataONE implementation shows that there are significant costs to providing robust support for provenance, including back-end software development, creation of appropriate and flexible representational structures, and the development of user interfaces to make provenance data accessible to repository users [30]. In addition, DataONE has extended multiple end-user tools (R and MATLAB) to enable scientists to capture provenance data automatically during their routine work. However, while the infrastructure is now largely in place, will scientists begin to utilize these tools in the way intended? DataONE personnel are now working on this bootstrapping problem.

Uncertainty regarding the provenance of physical archival materials can be a problem for archivists and can, in certain cases, result in widely reported controversies [31]. The author's research program seeks to provide answers to questions about the utility of provenance data in varying contexts across libraries, archives, and museums, as noted above. Avenues of investigation include interviews of archivists, curators, developers of digital asset management systems, and staff at institutions where some of these digital asset management systems have been implemented. Areas of inquiry in the interviews include the degree of need for stable digital representation of provenance in digital asset management systems, issues of trust and information / metadata quality, and how provenance data is collected, organized, and transferred forward into successor systems.

ACKNOWLEDGMENT

This work was supported by Data Observation Network for Earth (DataONE), National Science Foundation award #0830955 under a Cooperative Agreement.

REFERENCES

[1] R. Erway (2013). "Starting the conversation: University-wide research data management policy," Educause Review, Retrieved April 25, 2016, from http://er.educause.edu/articles/2013/12/starting-the-conversation-universitywide-research-data-management-policy

[2] P. T. Doran and M. K. Zimmerman (2009). "Examining the scientific consensus on climate change," Eos Transactions American Geophysical Union 90(3), pp. 22–23, doi:10.1029/2009EO030002.

[3] K. G. Akers, F. C. Sferdean, N. H. Nicholls, and J. A. Green (2014). "Building support for research data management: Biographies of eight research universities," International Journal of Digital Curation, 9(2), pp. 171-191.

[4] C. Tenopir, D. Hughes, S. Allard, M. Frame, B. Birch, L. Baird, R. J. Sandusky, M. Langseth, and A. Lundeen (2015). "Research data services in academic libraries: Data intensive roles for the future?" Journal of eScience Librarianship, 4(2). http://dx.doi.org/10.7191/jeslib.2015.1085.

[5] C. Tenopir, R. J. Sandusky, S. Allard, and B. Birch (2013). "Academic librarians and data research services: Preparation and attitudes," IFLA Journal, 39(1), pp. 70-78. doi: 10.1177/0340035212473089.

[6] C. Tenopir, R. J. Sandusky, S. Allard, and B. Birch (2014). "Research data management services in academic research libraries and perceptions of librarians," Library & Information Science Research, 36(2), pp. 84-90. doi: 10.1016/j.lisr.2013.11.003.

[7] Y. Gil and S. Miles (2013). PROV Model Primer: W3C Working Group Note. Accessed June 8, 2015 at http://www.w3.org/TR/prov-primer/.

[8] C. T. Silva, and J. E. Tohline (2008). "Computational provenance," Computing in Science Engineering, 10(3), pp. 9–10. http://doi.org/10.1109/MCSE.2008.71.

[9] provenance, n. (n.d.). OED Online. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/153408.

[10] A. J. Gilliland-Swetland (2000). Enduring paradigm, new opportunities: The value of the archival perspective in the digital environment. Council on Library and Information Resources, February 2000. Accessed June 8, 2015 at http://www.clir.org/pubs/abstract/pub89abst.html.

[11] provenance | Society of American Archivists. (n.d.). Retrieved April 29, 2016, from http://www2.archivists.org/glossary/terms/p/provenance#.VyOSEPkrJhE.

[12] C. M. Dollar (1992). Archival Theory and Information Technologies: The Impact of Information Technologies on Archival Principles and Methods. Macerata, Italy: Publications of the University of Macerata.

[13] custodial history | Society of American Archivists (n.d.) Retrieved April 29, 2016, from http://www2.archivists.org/glossary/terms/c/custodial-history#.VyOTefkrJhE

[14] P. Buneman, S. Khanna, and T. Wang-Chiew (2001). "Why and where: A characterization of data provenance," in Database Theory — ICDT 200, J. V. den Bussche and V. Vianu, Eds. Berlin: Springer, pp. 316–330.

[15] J. Freire, D. Koop, E. Santos, and C. T. Silva (2008). "Provenance for computational tasks: A survey," Computing in Science Engineering, 10(3), pp. 11–21. http://doi.org/10.1109/MCSE.2008.79.

[16] Graph theory. In Wikipedia, the free encyclopedia. Retrieved May 12, 2016, from https://en.wikipedia.org/w/index.php?title=Graph_theory&oldid=719689628.

[17] W. K. Michener, S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, and D. A. Vieglais (2012). "Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences," Ecological Informatics, 11, pp. 5-15, http://dx.doi.org/10.1016/j.ecoinf.2011.08.007.

[18] Immutability of Content in DataONE — v2.0-beta (n.d.) Retrieved May 15, 2016, from http://jenkins-1.dataone.org/jenkins/job/API%20Documentation%20-%20trunk/ws/api-documentation/build/html/design/ContentImmutability.html?highlight=provenance.

[19] B. Gunia and R. J. Sandusky (2010). "Designing metadata for long-term data preservation: DataONE case study," Proceedings of the American Society for Information Science and Technology, 47(1), n.p. http://doi.org/10.1002/meet.14504701435.

[20] PREMIS Editorial Committee (2008). Data Dictionary for Preservation Metadata: PREMIS version 2.0. Retrieved May,13 2016, from http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

[21] Scientific Workflows and Provenance Working Group | DataONE. (n.d.). Retrieved May 13, 2016, from https://www.dataone.org/working_groups/scientific-workflows-and-provenance-working-group.

[22] P. Missier, K. Belhajjame, and J. Cheney (2013). "The W3C PROV family of specifications for modelling provenance metadata," Proceedings of the 16th International Conference on Extending Database Technology. New York, NY, USA: ACM, pp. 773–776. http://doi.org/10.1145/2452376.2452478.

[23] PROV-DM: The PROV Data Model. (n.d.). Draft May 1, 2015. Retrieved May 13, 2016, from https://www.w3.org/TR/prov-dm/.

[24] The ProvONE Data Model for Scientific Workflow Provenance. (n.d.). Retrieved May 15, 2016, from http://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html. .

[25] Open Archives Initiative Protocol - Object Reuse and Exchange. (n.d.). Retrieved May 15, 2016, from http://www.openarchives.org/ore/.

[26] M. Carls (2015). Hydrocarbon database, Gulf of Alaska. MN Demo 2. Retrieved May 16, 2016 from https://search.test.dataone.org/#view/urn:uuid:bf71c38b-22b2-469e-8983-734ec0ab19c.

[27] Exxon Valdez oil spill (2016, May 14) In Wikipedia, the free encyclopedia. Retrieved May 14, 2016 from https://en.wikipedia.org/w/index.php?title=Exxon_Valdez_oil_spill&oldid=720201904.

[28] DataONEorg/matlab-dataone (n.d.) Retrieved May 15, 2016, from https://github.com/DataONEorg/matlab-dataone

[29] GitHub - NCEAS/recordr: Provenance tracking for R (n.d.) Retrieved May 15, 2016, from https://github.com/NCEAS/recordr.

[30] DataONEorg/sem-prov-design (n.d.) Retrieved May 12, 2016, from https://github.com/DataONEorg/sem-prov-design

[31] V. Harris and K. Stine (2011) "Politically charged records: a case study with recommendations for providing access to a challenging collection," The American Archivist: 74(2), pp. 633-651. http://dx.doi.org/10.17723/aarc.74.2.f252r28174251525