Provenance and data-intensive science

Michael Day[1]

UKOLN, University of Bath, Bath BA2 7AY, United Kingdom
m.day@ukoln.ac.uk

Abstract. [tba]

1. Introduction

The second edition of the *Oxford English Dictionary* (1989) defines provenance as "the fact of coming from some particular source or quarter; origin; derivation," noting that the term is derived from the Latin *provenire* by way of the French *provenant*. [The provenance of scientific ideas and data has always been important in science], However, the dependence of current scientific research means that there . Partly this is purely a matter of scale. Science is becoming increasingly dependent on the generation and reuse of massive amounts of data, a trend sometimes known as data-intensive science. [data deluge]. Examples include data being transmitted from earth observation instruments based on satellites, space telescopes, particle accelerators, Bioinformatics data

Distributed

Modern scientific enquiry is increasingly dependent on data, intensive ... data from space telescopes, particle accelerators, genomics, high-throughput combinatorial chemistry,

Primary scientific data is [...]. Data generation scaling up, petabytes of data expected from the Large Hadron Collider (LHC) under construction at CERN. Hey and Trefethen have talked about a data deluge, need to manage this task. In the UK the Joint Information Systems Committee and the e-Science Core programme funded a DCC to look at [...]

Research topics include the unique identification and citation of data and

This report will review the its importance for the reuse of scientific data in It will then look at the concept of provenance as it is applied in the contexts of science, art history and museums and archives. A final section will [...]

[1] Working draft, v. 0.1, 25 January 2005

2. Provenance and curation

Curation is the name given to the

2.1 Data-intensive science

One motivation of the eBank project is the massive growth in the generation and curation of scientific data typified by the advent of e-science and Grid computing. Much scientific research is now data intensive. There are many examples of scientific instruments and experiments that are (or will soon be) generating vast amounts of data.

- *High-energy physics*. Perhaps the most cited example of massive data generation is the Large Hadron Collider (LHC) currently under construction at CERN, the European Organization for Nuclear Research, near Geneva (Myers, 2003). From around 2007, experiments on this accelerator will be used to explore fundamental mechanisms in particle physics, including the search for the Higgs boson and supersymmetic particles (e.g., Renton, 2004; Khalil, 2003). It is estimated that once LHC is operational, these experiments will generate in the region of 12-14 petabytes of data per year (http://lcg.web.cern.ch/lgc/). This mass of data will be stored and managed across multiple international sites through the LHC Computing Grid (LCG) project. Future particle accelerators like the proposed International Linear Collider (ILC) are likely to have even more intensive data requirements.
- *Astronomy*. Other examples of Examples here might include digital sky surveys like the Sloan Digital Sky
- *Earth observation*. Satellites can also be used to . NASA's Earth Observing System (EOS) , EOS Data and Information Sysetm (EOSDIS) the outputs of earth observation satellites EOSDIS and [Szalay?].
- *Bioinformatics*. genomics
- *Materials science*. the introduction of combinational chemistry and high-throughput experimentation has resulted in an explosion of experimental data that requires organisation, administration, storage and evaluation (Adams & Schubert, 2004). The growing dependence of science on vast amounts of data have Hey and Trefethen (2003) have described this as a 'data deluge.'

Work is already proceeding on. What the data deluge means in practice is that only a very small percentage of the data generated by experiments or observation is published in or referenced by the scientific literature. In addition, the traditional, journal-based, scientific literature often only provides indirect access to data [??]. Access to data is important in science in order to test and replicate results.

2.x Fighting scientific misconduct

It can also be important in helping to combat scientific misconduct e.g. the recent case of the fraudulent manipulation and misrepresentation of data by the Bell Labs physicist Jan Hendrik Schön (e.g., Durrani, 2002).

3. Provenance

Provenance This section will introduce some of the more specialised understandings of provenance that have been developed by art historians, museum curators and archivists. [rewrite]
[stuff from TFADI report, informs OAIS model Provenance information,, important in discussions of artworks or archaeological artifacts, also elevated as a principle in archives.

3.1 Approaches to data provenance

We have already noted the increased ... focus on derivation, lineage, ... [tba, will focus on Buneman's studies of databases??, myGrid stuff, specifics to chemistry

Data provenan

Existing studies of data provenance

Provenance in the OAIS Reference Model

Provenance is one aspect of the Preservation Description Information defined by the Reference Model for an Open Archival Information System (OAIS).

Chimera (Ian Foster), myGrid (Goble),

3. Provenance in museums and archives

Research scientists are not the only ones who find the concept of provenance useful. Some cultural heritage organisations routinely record the origin of certain classes of object. Provenance is of particular importance to museums, where it can be used to help understand the origin, transmission and chain of ownership of the objects that they collect. Provenance information, for example, might record the location and date that a particular object was found. This is of particular importance for archaeological artefacts, geological specimens or collections of biodiversity information, where find data can itself be combined and reanalysed to help develop a deeper understanding, e.g. of archaeological contexts, geomorphology or ecosystems. In some cases, there may be an additional attempt to trace objects back to their ultimate origins, e.g. provenance research in archaeology relates to attempting to discover where raw materials were mined or an object manufactured. Methods like neutron activation analysis (NAA) are sometimes used to trace artefacts back from their find spot to their place of origin, e.g. for ceramics, glass or coins (e.g., Glascock & Neff, 2003). Similar

principles apply in geology, e.g. with regard to the composition of sandstones or conglomerates, and in forestry. [chain of custody, origin in collections, relate to indovidial scientists, experiments, expeditions].

[tba] Important part of the information associated with artworks or archaeological artefacts. Link with science, important for collections of biodiversity data (palaeontology, etc), in sense of basic data on where found, date and collector. Archaeology, test AHDS ... Art history, having impact on value, known chain of ownership reduces chance of fraud or may help identify stolen items ... SPECTRUM standard

3.x Provenance in art history

The OED's definitions of provenance include one related specifically to the history or pedigree of works of art, manuscripts, rare books, etc. as, "a record of the ultimate derivation and passage of an item through its various owners." This is how provenance has been commonly understood in the art world. One of the authorities cited by the OED (Barron, 1967) defines the term in an inclusive way.

> *Provenance*, a history or pedigree of a painting, the establishment of the identity of successive owners since its execution. Also included would be all published documents, catalogues, and journals that contain references to the painting, along with reproductions, exhibitions, and sales records, as well as correspondence, especially of the artist, in which mention of it may be made.

The study of provenance (or the history of collecting) has become an important topic of research in the art history domain. Scholars are interested in investigating chains of ownership as a means of providing information on the history of individual artworks and of collecting. Museums are interested in provenance as a means of authenticating artworks or establishing its legal ownership, e.g. part of the ongoing debate over the custody of the Elgin Marbles by the British Museum centres on this latter issue (ref?). The importance of provenance means that museums and art history scholars have developed resources to support its study. For example, the Getty Research Institute have created and maintain Provenance Databases (http://piweb.getty.edu/) which can be used to trace the chain of the known ownership of art works, with information taken from the catalogues of public collections, sales documents, and other documentary records. The level of detail included varies. For example the record for Titian's *Bacchus and Ariadne* contains brief information on 20 separate owners of the work, from Alfonso I d'Este, who initially commissioned the work for the Camerino d'Alabastro in Ferrara, through the Aldobrandi family, who confiscated it and held it in Rome for around 200 years, through its sale and arrival in London (in private ownership) at the turn of the 19th century, to it finally entering the National Gallery's collections in 1826. Other entries in the database are much less detailed.

Provenance research sometimes has political significance. For example, the National Socialist regime in Germany indulged in the confiscation of artworks from Jews in an officially sanctioned policy of *Kunstraub* (art theft); first in post-*Anschluss* Austria, then in Germany and the other occupied territories. After the invasion of Poland in 1939, this expanded into a more thoroughgoing expropriation of cultural artefacts. Much of this looted art ended up in German galleries, state or party institu-

tions, or in the personal collections of the Nazi elite (Petropoulos, 1996). Other works simply vanished. Partly as a response, the Soviet Union also removed major collections of cultural artefacts from Germany at the end of the war as a form of 'compensatory restitution' (Piotrovsky, 2004). Some of these, including the Pergamon Altar, were returned to the German Democratic Republic in the 1950s, but others still remain in Russian custody. Major efforts were made by the Allies after the war to collect cultural artefacts looted by the Nazis and return them to their original owners (Nicholas, 1994). Inevitably, however, the provenance of some items could not be identified while other artefacts ended up on the international art market and have since been dispersed into public and private collections. Databases of Nazi-era art exist to help people looking for lost artefacts and to help institutions fulfil their obligations to support the identification of artefacts stolen by the Nazis. For example, the Nazi-Era Provenance Internet Portal provides a registry of objects in US museum collections that changed hands in Europe between 1933 and 1945 (http://www.nepip.org/). In the UK, the National Gallery and the National Museum Directors' Conference both maintain information on artworks in public collections that have an incomplete provenance for the Nazi-era.

3.x The principle of provenance in archival science

While art historians, archaeologists, and others consider provenance to be an important factor in recording the history and context of objects, the archival profession has elevated the concept to a general principle. For archivists, provenance and the related concept of 'original order' informs almost every part of archival theory and practice. Provenance and original order together make up the wider archival principle of *respect des fonds*. It is the insight of archivists that the authenticity and integrity of records depends, at least in part, in tracing their origin and past history.

European archivists first elaborated the principles of *respect des fonds* and provenance in the 19th century. *Respect des fonds* was first formulated in a French ministerial circular of 1841 written by Natalys de Wailly (Ogilvie, 2002). Up until that point, archives had often been sorted according to subject, date or place, on occasion according to elaborate methodological schemes. This, after all, was the age of the great scientific classification schemes of Cuvier, Linnaeus and Berzelius (Duchein, 1977). In place of this, de Wailly proposed that a better way of organising archives would be to unite all documents that come from a particular organisation or individual, i.e. to classify archives by *fonds*. Later in the 19th century, German archivists like Max Lehmann developed a more thoroughgoing principle of provenance (*Provenienzprincip*), which from the 1880s was applied to the arrangement of the Prussian Privy State Archives. Posner (1967) explains that the motive for adopting the principle of provenance as was the unsatisfactory nature of the then existing subject-based arrangement of the Prussian Archives, which made finding records an extremely cumbersome procedure. The provenance-based system of arrangement also corresponded well with the new source-based historical approaches of Friedrich von Ranke and his contemporaries. European principles and practices were later codified in the Dutch *Manual for the classification and description of archives* (Muller, Feith & Fruin, 2000), first

published in 1898. Nesmith (1993) has described the 19th century discovery of the contextual approach to archival administration in Europe as "the most important intellectual development in the history of the archival profession." Cook (1993, p. 26) says that when archivists adhere to the principles of provenance and original order, "the evidential character of archives is protected, whereby the records inherently reflect the functions, programmes and activities of the person or institution that created them, and the transactional processes by which that actual creation took place."

Provenance-based thinking gradually permeated archival thinking across the whole world. A provenance-based approach to arranging archives had also been adopted by the Public Record Office in the UK, although a narrow focus on the 'archive group' (or highest level of administrative structure) resulted in the reassertion of the importance of provenance and original order by Sir Hilary Jenkinson and others in the early 20th century (Roper, 1992). [In his *Manual of archival administration*, Jenkinson (1965) himself proposed the "arrangement or exposition of the administrative object which the archive originally served" [...] p. 97-98, 101, 103. [c.f. also Levene, EHR, 101 (1986) 20-41.] Something on the US 'record group,' Australian views Peter Scott series system [rewrite paragraph...]

Towards the end of the 20th century, partly in response to the challenge of electronic records, some practical aspects of the application of the principle of provenance in archival science became subject to a reappraisal. For example, David Bearman and Richard Lytle argued that the 'record group' concept in the US assumes a hierarchical view of organisations and records that is at odds with the more complex and fluid reality of modern organisations (Bearman, 1992). Instead they note the continued importance of provenance information as access points for retrieval and emphasise the importance of the form of material and function. They say that because "archival records are the consequences of activities defined by organizational functions, such a vocabulary can be a powerful indexing language to point to the content of archival holdings, without need for actual examination of the materials themselves or for detailed subject indexing" (Bearman & Lytle, 1985, p. 22). They also proposed the establishment of provenance authority records that should be maintained separately from record descriptions.

In Canada, Terry Cook has taken a similar approach to the archival fonds as a conceptual rather than a physical entity. Building on Bearman and Lytle's analysis of modern organisations as non-hierarchical, fluid entities, Cook has also argued that the development of electronic records has highlighted weaknesses in the traditional understanding of the archival fonds.

> As businesses and governments ... adopt new information models based on corporate data planning and data resource management, the idea of a record physically belonging in one place or even in one system is crumbling before the new *conceptual* paradigms, where 'creatorship' is a fluid process of manipulating information from many sources in a myriad of ways rather than something leading to a static, fixed, *physical* product (Cook, 1992, p. 63).

In place of the traditional emphasis on physical arrangement in the record group or series, Cook suggests that the archival fonds should be better regarded as an abstract concept and that providing managed links between record descriptions and authority control entries for creating agencies might be a better way of recording the ever-

changing provenance of records. [Influence on archival description schemes, Canadian RAD, MAD, ISAD(G), EAD, Guercio].

x.x Lessons for science

As we have seen, the principle of provenance in archives , lessons for scientific data might be emphasis on accurate modelling of data and the organisations that create it, whether this be universities or research institutes, departments, research teams, principal investigators. look at CCLRC metadata schema to see if this is modelled accurately.

4. Citation and unique identification

5. Provenance and the eBank project

[Rewrite, need less on project itself, more on workflow and processes, the metadata and identifiers used to facilitate it]
The eBank UK project has been funded by the Joint Information Systems Committee (JISC) and is led by UKOLN at the University of Bath in partnership with the Universities of Southampton and Manchester (Heery, *et al*., 2004). The project is investigating the role of aggregator services in linking metadata describing scientific papers to datasets made available from repositories. The primary subject focus of the project is chemistry and eBank is working with the University of Southampton's project Comb*e*chem, which is funded by the Engineering and Physical Sciences Research Council (EPSRC) as an e-science testbed (Frey, *et al*., 2003). Initially, an eBank demonstrator service is being developed within the sub-domain of crystallography, but the project will also assess the general feasibility of the approach within other parts of chemistry and other scientific disciplines.

The eBank project is advocating a 'publication at source' philosophy based on open access principles [ref]. It is envisaged that research teams could routinely deposit datasets in institutional repositories as part of the data creation workflow. This would not only enable the subsequent linking of these datasets to peer-reviewed papers or results data published in specialised databases, but also enable certain aspects of the data creation and enhancement process to be recorded and (if necessary) repeated. In addition, the links thus created would help to record aspects of the provenance of datasets, ultimately linking back (for example) to records in laboratory notebooks.

For the development of its demonstrator service, the eBank project decided to focus on crystallography, as this sub-discipline has a well-defined data creation workflow and a tradition of sharing results data in an internationally accepted standard, the Crystallographic Information File (CIF) adopted by the International Union of Crystallography (Hall, Allen & Brown, 1991; Brown & McMahon, 2002). In addition,

secondary services like the Cambridge Structural Database (CSD) provides facilities for the acquisition, storage, validation, retrieval, analysis and visualisation of small-molecule crystal structures, which are mostly available in CIF format (Allen, 2002). Many crystallographic journals encourage or mandate the submission of CIF data and the CSD acts as an official data depository on behalf of a number of these.

The architecture of the eBank demonstrator is based on metadata harvesting and uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This protocol is based on defining the interactions between two classes of service. OAI 'data providers' act as repositories of resources and make selected metadata about those resources available so that they can be harvested by one or more 'service providers.' In turn, service providers can harvest metadata from multiple data providers and provide a single interface that aggregates them into a single virtual collection. [Something on OAI implementation, e-prints repositories, images, etc. The OAI-PMH has been implemented widely by subject based and institutional repositories, ePrints UK]

For the eBank UK demonstrator, a test data provider repository of crystallography datasets has been created at the University of Southampton. It is envisaged that Crystallographers will deposit datasets into this repository at suitable stages in the data creation and processing workflow. This repository uses a specially enhanced version of the EPrints.org software developed at the University of Southampton (http://epints.soton.ac.uk/) and metadata about the datasets can be entered manually or generated automatically. The local repository can be used

6. Conclusions

[tba]

Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council (MLA), the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from JISC, the European Union and other organisations. UKOLN also receives support from the University of Bath, where it is based.

References

Achard, F., Vaysseix, G., & Barillot, E. (2001). "XML, bioinformatics and data integration." *Bioinformatics*, 17, 115-125.

Adams, N., & Schubert, U. S. (2004). "From data to knowledge: chemical data management, data mining, and modeling in polymer science." *Journal of Combinational Chemistry*, 6, 12-23.

Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." *Acta Crystallographica*, B58, 380-388.

Augen, J. (2002). "The evolving role of information technology in the drug discovery process." *Drug Discovery Today*, 7, 315-323.

Barron, J. N. (1967). *The language of painting: an informal dictionary*. Cleveland, Ohio: World Publishing.

Bearman, D. (1992). "Diplomatics, Weberian bureaucracy, and the management of electronic records in Europe and America." *American Archivist*, [...]

Bearman, D., & Lytle, R. H. (1985). "The power of the principle of provenance." *Archivaria*, 21, 14-27.

Brown, I. D., & McMahon, B. (2002). "CIF: the computer language of crystallography." *Acta Crystallographica*, B58, 317-324

Buneman, P., Khanna, S., & Tan, W. -C. (2000). "Data provenance: some basic issues." In: Kapoor, S., & Prasad, S. (eds.), *FST TCS 2000: foundations of software technology and theoretical computer science, 20th conference, New Delhi, India, December 13-15, 2000*. (Lecture Notes in Computer Science, 1974). Berlin: Springer-Verlag, 87-93. Retrieved May 17, 2004 from: http://db.cis.upenn.edu/DL/fsttcs.pdf

Buneman, P., Khanna, S., & Tan, W. -C. (2001). "Why and where: a characterization of data provenance." In: Bussche, J. van den, & Vianu, V. (eds.), *Database theory - ICDT 2001: 8th international conference, London, UK, January 4-6, 2001*. (Lecture Notes in Computer Science, 1973). Berlin: Springer-Verlag, 316-330. Retrieved May 17, 2004 from: http://db.cis.upenn.edu/DL/whywhere.pdf

Buneman, P., Khanna, S., Tajima, K., & Tan, W. -C. (2004). "Archiving scientific data." *ACM Transactions on Database Systems*, 29, 2-42.

Carugo, O., & Pongor, S. (2002). "The evolution of structural databases." *Trends in Biotechnology*, 20, 498-501.

Cook, T. (1992). "The concept of the archival fonds: theory, description, and provenance in the post-custodial era." In: Eastwood, T., (ed.), *The archival fonds: from theory to practice = Le fonds d'archives: de la théorie à la pratique*. Ottawa: Bureau of Canadian Archivists, Planning Committee on Descriptive Standards, 31-85.

Cook, T. (1993). "The concept of the archival fonds in the post-custodial era: theory, problems and solutions." *Archivaria*, 35, 24-37.

Cook, T. (2001). "Fashionable nonsense or professional rebirth: postmodernism and the practice of archives." Archivaria, 51, 14-35.

Duchein, M. (1977). "Les respect des fonds en archivistique: principes théoriques et problèmes pratiques." *Gazette des Archives*, 97, 89-114.

Duff, W. M., & Harris, V. (2002). "Stories and names: archival description as narrating records and constructing meanings." *Archival Science*, 2, 263-285.

Durrani, M. (2002). "Misconduct strikes the heart of physics." *Physics World*, x(11), November, 6-7.

Foster, I., Vöckler, J., Wilde, M., & Zhao, Y. (2002). "Chimera: a virtual data system for representing, querying, and automating data derivation." In: Kennedy, J. (ed.), *Proceedings 14th International Conference on Scientific and Statistical Database Management, 24th-26th July 2002, Edinburgh, Scotland*. Los Alamitos, Calif.: IEEE Computer Society, 37-46. Retrieved May 17, 2004 from: http://www.globus.org/research/papers/VDS02.pdf

Frey, J. G., Bradley, M., Essex, J. W., Hursthouse, M. B., Lewis, S. M., Luck, M. M., Moreau, L. A. V. M., De Roure, D. C., Surridge, M., & Welsh, A. H. (2003). "Combinatorial chemistry and the Grid." In Berman, F., Fox, G., & Hey, A. J. G. (eds.), *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, 945-962.

Frey, J. G., De Roure, D., & Carr, L. (2002). "Publication at source: scientific communication from a publication web to a data grid." Euroweb 2002 - the Web and the GRID: from e-science to e-business [tba].

Frishman, D., Kaps, A., & Mewes, H. W. (2002). "Online genomics facilities in the new millennium." *Pharmacogenomics*, 3, 265-271.

Gilliland-Swetland, A. J. (2000). *Enduring paradigm, new opportunities: the value of the archival perspective in the digital environment.* Washington, D.C.: Council on Library and Information Resources.

Glascock, M. D., & Neff, H. (2003). "Neutron activation analysis and provenance research in archaeology." *Measurement Science and Technology*, 14, 1516-1526.

Gray, J., & Szalay, A. (2001). "The world-wide telescope." *Science*, 293, 2037-2040.

Greenwood, M., Goble, C., Stevens, R., Zhao, J., Addis, M., Marvin, D., Moreau, L., & Oinn, T. (2003). "Provenance of e-science experiments - experience from bioinformatics." UK e-Science All Hands Meeting, Nottingham, UK, 2-4 September 2003. Retrieved May 17, 2004 from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/047.pdf

Guercio, M. (1994). "Archival theory and the principle of provenance for current records." In: *The principle of provenance: report from the First Stockholm Conference on Archival Theory and the Principle of Provenance 2-3 September 1993*. (Skrifter utgivna av Svenska Riksarkivet, 10). Stockholm: Swedish National Archives, 1994, 75-86.

Guler, S., Eberhart, A., & Rojas, I. (2003). "Web-based exchange of biochemical information." *Bioinformatics*, 19, 1730-1731

Hall, S. R., Allen, F. H., & Brown, I. D. (1991). "The Crystallographic Information File (CIF): a new standard archive file for crystallography." *Acta Crystallographica*, A47, 655-685.

Heery, R., Duke, M., Day, M., Lyon, L., Hursthouse, M. B., Frey, J. G., Coles, S. J., Gutteridge, C., & Carr, L. A. (2004). "Integrating research data into the publication workflow: the eBank UK experience." [details tba] Frascati, Italy, October 2004.

Hey, A. J. G. & Trefethen, A. (2003). "The data deluge [...]." In Berman, F., Fox, G., & Hey, A. J. G. (eds.), *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, [pp].

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H, *et al*. (2003). "The systems biology markup language (SMBL): a medium for representation and exchange of biochemical network models." *Bioinformatics*, 19, 524-531.

Jenkinson, H. (1937). *A manual of archive administration*, 2nd ed. London: Lund, Humphries, 1965 (reprint).

Khalil, S. (2003). "Search for supersymmetry at LHC." *Contemporary Physics*, 44(3), 193-201.

Lyon, L. (2003). "eBank UK: building the links between research data, scholarly communication and learning." *Ariadne*, 36. Retrieved May 17, 2004 from: http://www.ariadne.ac.uk/issue36/lyon/

Matthews, B., & Sufi, S. (2001). "The CLRC Scientific Metadata Model - Version 1." Technical Report DL TR 02001. Warrington: CLRC Daresbury Laboratory. Retrieved May 17, 2004 from: http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf

Muller, S., Feith, G. A., & Fruin, R. (1898). *Handleinding voor het Ordenen en Beschrijven van Archiven*. Groningen: [?].

Murray-Rust, P. (1998). "The globalisation of crystallographic knowledge." *Acta Crystallographica*, D54, 1065-1070

Murray-Rust, P., & Rzepa, H. S. (2002). "STMML: a markup language for scientific, technical and medical publishing." *Data Science Journal*, 1, 1-65. Retrieved May 17, 2004 from: http://journals.eecs.qub.ac.uk/codata/Journal/contents/1_2/1_2pdfs/ds121.pdf

Murray-Rust, P., & Rzepa, H. S. (2003). "Chemical markup, XML, and the World Wide Web, 4: CML Schema." *Journal of Chemical Information and Computer Sciences*, 43, 757-772.

Murray-Rust, P., Glen, R. C., Rzepa, H. S., Stewart, J. J. P., Townsend, J. A., Willighagen, E. L., & Zhang, Y. (2003). "A semantic GRID for molecular science." UK e-Science All Hands Meeting, Nottingham, UK, 2-4 September 2003. Retrieved May 17, 2004 from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/157.pdf

Myers, J. D., Allison, T. C., Bittner, S., Didier, B., Frenklach, M., Green, W. H., et al. (2004). "A collaborative informatics infrastructure for multi-scale science." Challenges of Large Applications in Distributed Environments (CLADE) Workshop, Honolulu, Hawaii, USA, 7 June 2004. Retrieved May 17, 2004, from: http://scidac.ca.sandia.gov/Get/File-886/CLADE_2004_3_28.PNNL-SA-40934.pdf

Myers, J. D., Chappell, A. R., Elder, M., Geist, A., & Schwidder, J. (2003). "Reintegrating the research record." *IEEE Computing in Science & Engineering*, 5(3) 44-50. Retrieved May 17, 2004 from: http://collaboratory.emsl.pnl.gov/presentations/papers/reintegrating.html

Myers, J. D., Pancerella, C., Lansing, C., Schuchardt, K. L., & Didier, B. (2003). "Multi-scale science: supporting emerging practice with semantically derived provenance." ISWC 2003 Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Fl., USA, 20 October 2003. Retrieved May 17, 2004 from: http://scidac.ca.sandia.gov/Get/File-855/prov_1.pdf

Myers, S. (2003). "Particle accelerators: to the LHC and beyond." *Physics World*, 39-43.

Navarro, J. D., Niranjan, V., Peri, S., Jonnalagadda, C. K., & Pandey, A. (2003). "From biological databases to platforms for biomedical discovery." *Trends in Biotechnology*, 21, 263-268.

Nesmith, T. (1993). "Archival studies in English-speaking Canada and the North American rediscovery of provenance." In: Nesmith, T., (ed.), *Canadian archival studies and the rediscovery of provenance*. Metchuen, N.J.: Scarecrow Press, 1-28.

Nicholas, L. H. (1994). *The rape of Europa: the fate of Europe's treasures in the Third Reich and the Second World War*. London: Macmillan.

Ogilvie, D. (2002). "La genèse de la théorie du respect des fonds: le classement par matières des archives administratives par Natalys de Wailly." Archivi e storia nell'Europa del XIX secolo, Archivio di Stato di Firenze, Florence, Italy, 4-7 December 2002. Retrieved May 17, 2004 from: http://www.archiviodistato.firenze.it/atti/aes/ogilvie.pdf

Pancerella, C., Hewson, J., Koegler, W., Leahy, D., Lee, M., Rahn, L., *et al*. (2003). "Metadata in the Collaboratory for Multi-scale Chemical Science." DC-2003: the 2003 Dublin Core Conference, Seattle, Washington, USA, 27 September - 2 October 2003. Retrieved May 17, 2004 from: http://purl.oclc.org/dc2003/03pancerella.pdf

Petropoulos, J. (1996). *Art as politics in the Third Reich*. Chapel Hill, N.C.: University of North Carolina Press.

Piotrovsky, M. (2004). "On the question of the existence of a cultural and historical memory in contemporary Russia." Retrieved May 17, 2004 from the Hermitage Museum Web site: http://www.hermitagemuseum.org/html_En/02/hm2_6_0_6.html

Posner, E. (1967). "Max Lehmann and the genesis of the principle of provenance." In: Posner, E., *Archives and the public interest*. Washington, D.C.: Public Affairs Press, 36-44.

Renton, P. (2004). "Has the Higgs boson been discovered?" *Nature*, 428, 141-144.

Roper, M. (1992). "The development of the principles of provenance and respect for original order in the Public Record Office." In: Craig, B. L., (ed.), *The archival imagination: essays in honour of Hugh A. Taylor*. Ottawa: Association of Canadian Archivists, 134-153.

Schuchardt, K., Didier, B., & Black, G. (2002). "Ecce: a problem solving environment's evolution toward Grid services and a Web architecture." *Concurrency and Computation: Practice and Experience*, 14, 1221-1239. Retrieved May 17, 2004 from: http://ecce.emsl.pnl.gov/docs/publications/ecce_evolution.pdf

Stevens, R., Glover, K., Greenhalgh, C., Jennings, C., Pearce, S., Li, P., Radenkovic, M., & Wipat, A. (2003). "Performing *in silico* experiments on the Grid: a users perspective." UK e-Science All Hands Meeting, Nottingham, UK, 2-4 September 2003. Retrieved May 17, 2004 from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/010.pdf

Sufi, S., Matthews, B., & Kleese van Dam, K. (2003). "An interdisciplinary model for the representation of scientific studies and associated data holdings." UK e-Science All Hands Meeting, Nottingham, UK, 2-4 September 2003. Retrieved May 17, 2004 from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/020.pdf

Szomszor, M., & Moreau, L. (2003). "Recording and reasoning over data provenance in Web and Grid services." In: *On the move to meaningful Internet systems 2003: CoopIS, DOA, and ODBASE*. (Lecture Notes in Computer Science, 2888). Berlin: Springer-Verlag, 603-620. Retrieved May 17, 2004 from: http://www.ecs.soton.ac.uk/~lavm/

Valencia, A. (2002). "Search and retrieve." *EMBO Reports*, 3, 396-400.

Wang, L. C., Riethoven, J. J., & Robinson, A. (2002). "XEMBL: distributing EMBL data in XML format." *Bioinformatics*, 18, 1147-1148

Waugh, A., Gendron, P., Altman, R., Brown, J. W., Case, D., Gautheret, D., Harvey, S. C., Leontis, N., Westbrook, J., Westhof, E., Zuker, M., & Major, F. (2002). "RNAML: a standard syntax for exchanging RNA information." *RNA*, 8, 707-717.

Wroe, C., Goble, C., Greenwood, M., Lord, P., Miles, S., Papay, J., Payne, T., & Moreau, L. (2004). "Automating experiments using semantic data on a bioinformatics grid." *IEEE Intelligent Systems*, 19(1), 48-55.

Zhao, J., Goble, C., Greenwood, M., Wroe, C., & Stevens, R. (2003). "Annotating, linking and browsing provenance logs for e-science." ISWC 2003 Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, Fl., USA, 20 October 2003. Retrieved May 17, 2004 from: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-83/prov_2.pdf