Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences*

TRACEY P. LAURIAULT, BARBARA L. CRAIG,
D.R. FRASER TAYLOR, and PETER L. PULSIFER

RÉSUMÉ Les données scientifiques sont essentielles à la formation en sciences et à la prise de décision éclairée au sujet de la santé, de l'environnement et de l'économie. Les ensembles de données cumulatives aident à comprendre les tendances, les fréquences et les courants, et ils peuvent servir de base pour développer des prévisions. Cet article se penche sur la préservation des données scientifiques et des portails de données scientifiques d'un ensemble de domaines, en ciblant la qualité des données – surtout l'exactitude, la fiabilité et l'authenticité – et en examinant comment ces caractéristiques sont saisies par les métadonnées. Les auteurs donnent des définitions générales de ces concepts, dans des perspectives à la fois scientifiques et archivistiques. À partir d'une recension approfondie de la littérature sur le sujet (publications provenant d'organisations scientifiques nationales et internationales, d'organismes gouvernementaux et d'organismes de financement, ainsi que des observations empiriques d'un échantillon d'études de cas d'InterPARES 2 et de « General Study 10 » qui étudiaient 32 portails de données scientifiques), cet article examine sommairement la « représentation des connaissances » électronique (« *machine-base "knowledge representation" [KR]* ») et les répercussions possibles sur la préservation des données scientifiques, avec un accent particulier sur les ontologies formelles. Il présente aussi le concept de document dans le contexte d'un environnement Web 2.0, la rareté des archives sur les données scientifiques, et le fait que ce domaine ne figure pas souvent dans les priorités de financement. Les auteurs avancent que les archivistes devront travailler de près avec les scientifiques créateurs de données afin de comprendre leurs pratiques; que les portails de données sont des mécanismes dont les archivistes peuvent se servir pour parfaire leurs pratiques de préservation; et que ce

* This paper is the result of two SSHRC-funded research projects, InterPARES 2 and Cybercartography and the New Economy. The authors would like to acknowledge the support of the InterPARES 2 UBC students who assisted with the collection of data for General Case Study 10: Sherry Xie, Heather Dean, Cristina Miller, Brian Tremblanth, and Stephen Gage. In addition, we are grateful to Bonnie Mak for facilitating research activities between Carleton and UBC, and Randy Preston, Greg Kozak, and Yau Min Chong for their coordinative support, and finally Jean-Pascal Morghese for ensuring that the extensive documentation related to the Cybercartographic Atlas Case Study and General Study 10 were made accessible to all researchers on the InterPARES 2 project.

n'est pas la technologie qui empêche le progrès en ce qui concerne les données scientifiques. C'est plutôt le manque de ressources, de politiques, de classement par ordre de priorités, et de vision qui occasionne la perte de nos ressources scientifiques nationales.

ABSTRACT Scientific data are essential for training in science and informed decision-making regarding health, the environment, and the economy. Cumulative data sets assist with understanding trends, frequencies and patterns, and can form a baseline upon which we can develop predictions. This paper discusses the preservation of scientific data, providing an overview of the characteristics of scientific data and scientific-data portals from a variety of fields, with a focus on data quality, particularly accuracy, reliability and authenticity, and how these are captured in metadata. These concepts are broadly defined from both scientific and archival perspectives. Based on an extensive literature review of publications from national and international scientific organizations, government and research funding bodies, and empirical evidence from a selection of InterPARES 2 Case Studies and General Study 10, which investigated thirty-two scientific-data portals, the paper includes a brief examination of machine-base "knowledge representation" (KR) and the potential implications for the preservation of scientific data, with a particular focus on formal ontologies. The paper also discusses the concept of record in the context of Web 2.0 environments, the paucity of scientific data archives, and the lack of funding priorities in this area. It is argued that archivists will have to work closely with scientific-data creators to understand their practices, that data portals are mechanisms that archivists can use to extend their preservation practices, and that it is not technology that is impeding progress regarding the preservation of scientific data; it is a lack of funding, policy, prioritizing, and vision allowing our scientific national resources to be lost.

Introduction: The Rationale for the Preservation of Scientific Data

Scientific data represent both the organization and chaos of the natural world. They stimulate us to develop new concepts, theories, and models to make sense of the patterns they represent. The resulting abstractions are the product of scientific endeavour, the goal being to develop the formal and systematic ideas that constitute the understanding of relationships between causes and consequences and perhaps may enable prediction of future sequences of events. Because scientists transform data from the material world into ideas, the observations of objects and processes in the physical world are the stimuli for scientific thought. Data are thus the seeds of scientific thought.

National Research Council[1]

1 National Research Council [NRC], Commission on Physical Sciences, Mathematics, and Applications, *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources* (Washington, DC, 1995), http://www.nap.edu/catalog.php?record_id=4871 (accessed 23 August 2007), p. 10.

Knowledge and understanding are cumulative, thus the more complete the records of our world, the more we can gain from them. Climate change research, archeological data, space navigation engineering drawings, or census data are classic examples. It becomes imperative to preserve data precisely because "data are the primary building block of science. Furthermore, the meaningful access to reliable scientific data merits as much attention to acquisition of data as the preservation and archiving of scientific data."[2] Many collected data are derived from one-of-a-kind events: a volcanic explosion, a tsunami, ocean temperature at a time and place, an experiment, or pre-election polls. Cumulative sets of data can assist with understanding trends, frequencies and patterns, and can form a baseline upon which we can develop predictions; and the longer the record, the greater the confidence we can have in conclusions derived from them.[3] Cumulative data sets are essential to data models and simulations. Training in science is dependent upon the accessibility of existing scientific data.[4] With advancements in both scientific methods and computer technology, we can glean more from data than ever before, while preservation can provide the raw materials required for future unintended uses that come with greater advances. Furthermore, the cost of preserving data is almost always lower than the cost of recollecting them,[5] providing of course that recollection is possible! The animal, plant, and human space-flight data in the NASA Life Sciences Archive, for instance, can never be recollected.

The assembled record of observational or scientific data has dual value: it is simultaneously a history of events in the natural world and a record of human accomplishments. The history of the physical world is an essential part of our accumulating knowledge and the underlying data form a significant part of that heritage. They also portray a history of our scientific and technological development.[6]

Therefore scientific data in government, private, educational, and non-profit sector databases "constitute a critical national resource, one whose value increases as the data become more readily and broadly available."[7] It is cost effective to maximize the returns on these investments by preserving them and disseminating them widely.

2 CODATA Working Group on Archiving Scientific Data, http://www.nrf.ac.za/codata/ (accessed 9 January 2007).
3 NRC, *Preserving Scientific Data.*
4 CODATA Working Group on Archiving Scientific Data.
5 National Science Foundation [NSF], *Report of the National Science Board: Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (Arlington, 2005), http://www.nsf.gov/pubs/2005/nsb0540/ (accessed 23 August 2007).
6 NRC, *Preserving Scientific Data,* p. 11.
7 Ibid., p. 50.

There are also socio-economic reasons for the long-term archiving of scientific data, in addition to historical and scientific reasons. Scientific data have many industrial uses and other practical applications; deep-sea ocean topography and currents, for instance, are critical to the laying of submarine telecommunications cables. The costs of preserving and archiving data are relatively small in comparison with the costs of acquiring scientific records through observation. There is also an argument to be made that publicly-funded data should also be accessible data.

Unfortunately, the technological, institutional, and organizational issues related to the long-term preservation of data remain largely unresolved. The basic digital data upon which we depend to inform decisions on planning, health, emergency preparedness, industrial exploration, and research are rarely being effectively archived and preserved and, as a result, much is being lost, some permanently. John Roeder, a researcher on both InterPARES (IP) projects, discovered that one fifth of the data generated by the 1976 Viking exploration of Mars,[8] the entire 1960 US census,[9] and the works of nearly half of composers,[10] and one quarter of digital photographers[11] have been lost or threatened by technological obsolescence or inadequate preservation strategies. The Canada Land Inventory, for example, one of the world's first geomatics projects, had to be recovered at great expense to taxpayers since the recollection of those data was very cost prohibitive. It has been argued that "in archiving terms the last quarter of the 20th century has some similarities to the dark ages. Only fragments or written descriptions of the digital maps produced exist. The originals have disappeared or can no longer be accessed."[12] It has also been noted that "indeed digital technology is responsible for much of the loss, as storage technology has given a false sense of security against loss and obsolescence,"[13] and "an unprecedented firestorm is

8 Terry Cook, "It's Ten O'clock, Do You Know Where Your Data Are?" *Technology Review* 98 (1995), pp. 48–53 and Ross Harvey, "An Amnesiac Society? Keeping Digital Data for Use in the Future," *LIANZA 2000 Conference* (New Zealand, 2000).

9 Donald Waters and John Garrett, *Preserving Digital Information, Report of the Task Force on Archiving of Digital Information* (Washington, DC, 1996). http://www.rlg.org/en/page.php?Page_ID=114 (accessed 11 January 2007).

10 John Longton, *Record Keeping Practices of Composers Survey Report* (Vancouver, 2005), http://www.interpares.org/ip2/ip2_general_studies.cfm?study=27 (accessed 23 August 2007).

11 Jessica Bushey and Marta Braun, *Survey of Record-Keeping Practices of Photographers Using Digital Technology Final Report* (Vancouver, 2005).

12 D.R. Fraser Taylor, Tracey P. Lauriault, and Peter L. Pulsifer, "Preserving and Adding Value to Scientific Data: The Cybercartographic Atlas of Antarctica," presented at *PV2005; Ensuring Long-Term Preservation and Adding value to Scientific Technical Data* (Edinburgh, 2005).

13 David F. Strong and Peter B. Leach, *The Final Report of the National Consultation on Access to Scientific Data* (Ottawa, 2005), http://ncasrd-cnadrs.scitech.gc.ca/NCASRDReport_e.pdf (accessed 5 January 2007), p. 13.

incinerating Canada's digital research wealth."[14]

Technological innovations have expanded the possibilities for knowledge sharing and representation, and opened new pathways for knowledge discovery. However, the corollary of constant innovation is technological obsolescence. Rapid obsolescence of hardware, and changes in the capabilities and structures of associated software, pose a great challenge to scientists and scientific activities because change influences the continued accessibility of digital objects and may further affect any object's readability, intelligibility, and even its accuracy. Several reports of the United States National Research Council have been devoted to this problem. The energy and costs of dealing with this problem on an individual basis is considerable, and many scientists have called for the development of more generally available and tested strategies for long-term preservation of accurate and authentic digital data and data sets.

As archivists need to play a key role in these strategies, it is important for them to better understand the scientific context. To that end, this paper will advance the discussion of the preservation of scientific data, centering on the key InterPARES 2 Project (IP2) research themes of accuracy, reliability, authenticity, metadata, and the term record as they pertain to scientific data.[15] It is based on two main sources: an extensive literature review of publications from national and international scientific organizations, government, and research funding bodies; and the empirical evidence from a selection of IP2 Case Studies from the Scientific Focus (Appendix 1), two case studies from the Government Focus, which include geomatics data, and General Study 10 (GS10), which investigated thirty-two scientific-data portals (Appendix 2). The GS10 Survey was undertaken to collect information about the actual practices, standards, and protocols currently used by broadly defined existing data services, archives, repositories, or catalogues in the sciences. This paper provides a descriptive analysis of data gathered from the survey. Since the GS10 Survey was undertaken for exploratory purposes, the sample size from each scientific discipline is small, limiting cross-disciplinary analysis. The study does, however, provide a deeper understanding of practices in the natural and physical sciences as these pertain to portals, selected case studies, and

14 Social Science and Humanities Research Council (SSHRC), *Research Data Archiving Policy* (Ottawa, 2002), http://www.sshrc.ca/web/apply/policies/edata_e.asp (accessed 23 August 2007).

15 The International Research on Permanent Authentic Records in Electronic Systems Project (InterPARES) was led by the University of British Columbia and ran from 1999 to 2006. IP2 (2002–2006) delved into the issues of authenticity, reliability, and accuracy from the perspective of the entire life cycle of records, from creation to permanent preservation. It focused on records produced in complex digital environments in the course of artistic, scientific, and e-government activities. See http://www.interpares.org/ip2/ip2_index.cfm (accessed 6 September 2007).

their associated data. The paper also includes an exploratory review, which considers the importance of issues such as accuracy, reliability, and authenticity in the management of scientific data exchanged through portals. The choice of the portals considered was based on recommendations from IP2 researchers who were familiar with, and used, these in their own research work. The portals selected pertained to different communities of practice in sciences, such as health, astronomy, biology, engineering, statistics, genetics, geosciences, and ecology, to name a few.[16]

We begin by presenting an overview of the characteristics of scientific data and scientific data portals, then introduce concepts related to elements of data quality including accuracy, reliability, and authenticity. These concepts will be broadly defined from both scientific and archival perspectives, keeping in mind that scientific disciplines, sub-disciplines, and scientific institutions or communities adhere to their own specialized data quality measures, parameters, and practices. Science is a broad discipline, is thematically heterogeneous, and each field adheres to its own specific methodologies, tool, technologies, practices, and norms. We need only think of the great differences between particle physics, astronomy, meteorology, genomics, biology and geodesy, and their related subfields to see the divergence. Nonetheless, associated scientific concepts such as lineage,[17] objectivity and bias, error, and disclaimers are discussed, as is metadata, a critical aspect of scientific data that includes both lineage information and, in almost all cases where metadata exist, quality parameters; each scientific discipline has its own specificities and particular metadata practices. A brief examination of the emergence of machine-base "knowledge representation" (KR) and the potential implications of this movement for the preservation of scientific data is presented, focusing on a particular component of KR known as a formal ontology. These concepts are further refined using data gathered in selected Science Focus

16 For the purposes of this paper, the generic term "portal" refers to these services. The research was not intended to be exhaustive but to be an overview of the preservation structures in place or lack thereof in the examples surveyed by Lauriault and Craig in 2006. Details of the Case Studies, General Study 10, and Scientific Focus reports can be accessed from the IP2 website, http://www.interpares.org/ip2/ip2_case_studies.cfm (accessed 6 September 2007). Additional tables containing data gathered during the case studies can be found as a supplementary file attached to the electronic version of this article in *e-Archivaria*, and on the IP2 website at http://www.interpares.org/ip2/ip2_dissemination.cfm?proj=ip2&cat=pu-atcl-r .

17 In the field of geomatics, lineage means the history of the dataset, the dataset's pedigree as it changes form, its life cycle from collection to acquisition by a repository, through all the data set's stages of conversion, correction and transformations, and its parentage. See Derek G. Clarke and David M. Clark, "Lineage," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 13–30.

Case Studies and General Study 10. As the concept of record is integral to the archival community, we discuss this concept from the perspective of the creators of scientific records and contrast it with that of archivists. Finally, the state of selective digital-data archives initiatives is discussed, and the paper concludes with some general observations about how the challenges of archiving scientific information can be met.

Scientific Data

Scientific data are defined as "numerical quantities or other factual attributes generated by scientists and derived during the research process (through observations, experiments, calculations and analysis),"[18] and "numbers, images, video or audio streams, software and software versioning information, algorithms, equations, animations, or models/simulations."[19] In an archival definition, scientific data are "facts, ideas, or discrete pieces of information, especially when in the form originally collected and unanalyzed."[20]

Data can be acquired directly from experimentation in laboratories or from the physical world, or can be derived from evaluated published data. Distinctions are made between *raw* or *level 0 data* and *derived, refined, synthesized,* or *processed data*. Raw data are normally unprocessed, such as digital signals from a sensor or an instrument (e.g., unprocessed satellite images, thermometer readings); facts derived from a sample collected for an experiment (e.g., blood or ice cores); or facts collected by human observation (e.g., tree counts, bird sightings, or a census). Computations and data manipulations are related to research objectives and methodologies. Refined or processed data are raw data that have been manipulated, undergone computational modeling, been filtered through an algorithm, sorted into a table, or rendered into a map.

A data set can be compiled from observed data, or derived from other sources, which can be a version of the observed data, or derived from interpreted data. Interpreted data are data that cannot be returned to their original observations or measurements by reverse application of the data-processing steps and/or application of interpretative-transformation algorithms.[21]

Often, refined data are more comprehensible to non-specialist audiences such as archivists; however, it is important to note that each time data are manipulated they are further removed from their raw data source.

18 CODATA Working Group on Archiving Scientific Data.
19 NSF, *Report of the National Science Board*, p. 18.
20 Richard Pearce-Moses, *A Glossary of Archival and Records Terminology* (Chicago, 2007), http://www.archivists.org/glossary/ (accessed 23 August 2007).
21 Clarke and Clark, p. 16.

Computations are often irreversible; therefore it is important to preserve the original raw data. Intermediate data are also important. During the experimental process,

> ... researchers may often conduct variations of an experiment or collect data under a variety of circumstances and report only the results they think are the most interesting. Selected final data are routinely included in data collections, but quite often the intermediate data are either not archived or are inaccessible to other researchers. There is, however, the growing realization that intermediate data may be of use to other researchers.[22]

According to the National Science Foundation's 2005 *Report of the National Science Board*, data can be distinguished by how they were collected as (1) observational; (2) computational; or (3) experimental.[23] *Observational data*, such as direct observations of ocean temperature on a specific date, the film footage of an Antarctic ice sheet breakup, pre-election polls, or photographs of a meteorite cloud are historical recordings of particular events that cannot be replicated nor recollected. Many of the data portals examined in IP2's GS10 include observational data, such as the British Atmospheric Data Centre (IP2SF1) or the World Data Centre for Terrestrial Physics (IP2SF10) (See Appendix 2). *Computational data*, such as the results from a climate change model that includes comprehensive information about the model (e.g., descriptions of the hardware, software, original input data, and metadata) may be reproducible, and in this case, they may be less important than the model itself. Many of the so-called GRID portals examined in GS10 include large collaborative scientific models and their associated data: Earth Systems Grid (IP2SF36), San Diego Supercomputing Center (IP2SF22), or the Joint Centre for Structural Genomics (IP2SF21). *Experimental data* "such as measurements of patterns of gene expression, chemical reaction rates, or engine performance present a more complex picture."[24] These data may be reproducible; however, the cost of doing so is prohibitive, and it may not be possible to reproduce the same experimental conditions. Examples of portals that include experimental data include the National Institute of Health (IP2SF17), the National Cancer Registry (IP2SF16), the Functional Magnetic Resonance Imaging (fMRI) Data Centre (IP2SF7), the Cambridge Crystallographic Data Centre (IP2SF4), and the NASA Life Sciences Archive (IP2SF2).

22 NSF, *Report of the National Science Board,* p. 19.
23 Ibid.
24 Ibid., p. 19.

The National Oceanic and Atmospheric Administration (NOAA) (IP2SF25), which produces data for emergency preparedness and coastline management, organizes its data according to several criteria: 1) Original Data; 2) Synthesized Products; 3) Interpreted Products; 4) Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories; 5) Natural Resource Plans; 6) Experimental Products; and 7) Corporate and General Information. Since the data NOAA produces are very influential and the decisions based on them involve risk to human health and safety, each of these categories is very well described, undergoes rigorous error checks, has specific data-quality parameters, and adheres to the Office of Management and Budget (OMB) guidelines.[25]

None of the IP2 Case Studies outlined in Appendix 1 includes experimental data; all have observational data in a variety of forms, while five of the studies also include computational data or models. All but the Cybercartographic Atlas[26] have their data rendered or stored in a proprietary system. All the data are stored in a variety of databases or a searchable data portal. The Engineering project data (CS19),[27] the Alsace-Moselle Land Registry (CS18),[28] and some of the NASA Mars Global Surveyor data (CS08) in particular are inseparable from the systems within which they have been created and/or stored. This small sample of case studies and the surveyed data portals illustrate the complexities involved with the preservation of scientific data and the particularities of each scientific discipline's practice.

Portals

Data are considered to be the building blocks of scientific thought, and our understanding of the physical universe is built "on current and past studies in individual disciplines, by collecting and analyzing new types of data, and by

25 Federal agencies in the United States have responsibilities under the *Data Quality Act (Public Law 106- 554; H.R. 5658, Sec. 515)*. In accordance with this Act, the OMB has issued guidelines that "provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information ... disseminated by Federal agencies." OMB, *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies; Notice; Republication*, Federal Register, vol. 67, no. 36 (Washington, DC, Friday, 22 February 2002), www.whitehouse.gov/omb/fedreg/reproducible2.pdf (accessed 17 January 2007).

26 Tracey P. Lauriault and Yvette Hackett, *CS06 Cybercartographic Atlas of Antarctica* (Vancouver, 2005).

27 Kenneth Hawkins, *CS19 Authenticating Engineering Objects for Digital Preservation* (Vancouver, 2005).

28 Jean-François Blanchette, Françoise Banat-Berger, and Geneviève Shepherd, *CS18 Computerization of Alsace-Moselle's Land Registry* (Vancouver, 2004).

using past observations in entirely new ways not envisioned when the data were initially collected."[29] A step forward in science occurs when data are disseminated, especially to other scientists, who can examine them critically, duplicate them, and eventually improve upon them. Data portals also provide access to longitudinal data sets that facilitate knowledge creation; as CODATA notes, "scientific reasons for preserving and archiving data relate to the notion that knowledge creation is a cumulative process and that much more information and knowledge could be extracted from complete data records."[30] However, as IP2 researchers discovered, most of the data either stored or discovered in data portals are not archived. This is troublesome, as the data in these portals

> ... provide important baselines to track rates of change and computing the frequency of rare events. The re-analysis of existing data may lead to different conclusions. Thus, archived data allow for the formulation of new hypotheses and may unexpectedly change the relative importance of the data.[31]

Portals nonetheless provide a framework from which digital-data archivists can work and one which they can expand with policies, standards, and metadata. It is important for archivists to remember that scientists, research groups, funding agencies, scientific data organizations, and governments have already determined that the data are valuable enough to be paid for, collected, described, licensed, endorsed, organized, and disseminated.

Geomatics and science data are increasingly being discovered and accessed in data portals. Portals have a variety of names, such as data repositories, clearinghouses, catalogues, archives, geolibraries, and directories. In this paper, the term portal is used to encompass all of these. These digital data collections

> ... give researchers access to data from a variety of sources and enable them to integrate data across fields. The relative ease of sharing digital data – compared to data recorded on paper – allows researchers, students, and educators from different disciplines, institutions, and geographical locations to contribute to the research enterprise. It democratizes research by providing the opportunity for all who have access to these data collections to make a contribution.[32]

29 NRC, *Preserving Scientific Data,* p. 13.
30 CODATA Working Group on Archiving Scientific Data.
31 Ibid.
32 NSF, *Report of the National Science Board*, p. 14.

Portals, catalogues, and geolibraries make it possible for users to

> ... gather data germane to their own needs more readily, extract data from online and other electronic repositories, develop the information product they need, use the products for decision making, and contribute their locally gathered geoinformation and derived products to libraries or other repositories.[33]

Portals can provide all or some of the following services: search and retrieval of data, item descriptions, display services, data processing, the platform to share models and simulations, and the collection and maintenance of data. Much but not all of the data derived from portals are raw in nature and require the user to interpret, analyze, and/or manipulate them. The reasons for their creation are one-stop shopping, distributed responsibility over data sets, discoverability, and reduction in cost, since data are stored once and used many times.[34]

Clearinghouses,[35] directories, and catalogues are the technical embodiments of data-sharing policies. Individuals within organizations, research projects, or scientific collaborations register their data holdings in the portal via an on-line form organized according to a metadata standard, and then choose to make their data available for free, sale, viewing, or downloading.[36] Metadata standards "establish the terms and definitions to provide a consistent means to describe the quality and characteristics of geospatial data,"[37] and the ISO 19115 metadata standard[38] has become an international standard in the field of geomatics. Most of the portals examined in IP2's General Study 10 include either very detailed metadata or rudimentary header information that only contains lineage information.[39]

33 NRC, Spatial Information Resources, *Distributed Geolibraries* (Washington, DC, 1999), http://www.nap.edu/openbook.php?isbn=0309065402 (accessed 23 August 2007), p. 36.

34 Tracey P. Lauriault, "A Geospatial Data Infrastructure is an Infrastructure for Sustainable Development in East Timor" (Master's thesis, Carleton University, 2003).

35 Clearinghouses are a network of Internet portals that are interconnected by a common metadata cataloguing standard and by agencies that agree to make these portals accessible to each other.

36 Lauriault.

37 Nancy Tosta and Michael Domaratz, "The U.S. National Spatial Data Infrastructure," in *Geographic Information Research: Bridging the Atlantic*, ed. Massimo C. Craglia and Helen Couclelis (London, 1997), p. 22.

38 International Standards Organization, *Geographic information – Metadata* (2003), http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35&ICS2=240&ICS3=70 (accessed 10 January 2007).

39 See "Table 3: Metadata in Selected Science and Geomatics InterPARES 2 Case Studies" in the supplementary file attached to the electronic version of this article in *e-Archivaria*, or on the InterPARES 2 website at http://www.interpares.org/ip2/ip2_dissemination.cfm?proj=ip2&cat=pu-atcl-r .

The architecture of data portals varies. The National Research Council, Spatial Information Resources, Distributed Geolibraries report indicates that portals can be a single-enterprise sponsored portal (like a national library), a network of enterprises (like a federation of libraries) or a loose network connected by protocols (like the Web).[40] *Distributed data portals* have data sets described according to a given standard, and when a request is sent to them by a given site a search is executed by a search agent[41] to access or render the data into a map or some other form. The Cybercartographic Atlas of Antarctica, for example, adheres to interoperable Open Geospatial Consortium standards and specifications. When a user accesses a particular atlas module, a call is made to the British Antarctic Survey data portal in the United Kingdom, the data are accessed, rendered into a map in real time by the Atlas Framework in Ottawa, and delivered directly to the user's computer. Other examples of this type of portal are GRID portals, those which use web-mapping services, such as the British Antarctic Survey (IP2SF30) and the FMRI data centre (IP2SF7). A *Collection level catalog/portal* identifies a data custodian's holdings and uses them to direct searches.[42] Z39.50, a server-sharing standard that enables searching multiple data holdings, is an example of a collection portal mechanism, as is the Ocean Biogeographic Information System – Spatial Ecological Analysis of Megavertebrates Populations (IP2SF13). A *unified catalogue* exists in one place: data custodians submit metadata for each data set to a central site, which makes them available for searching, and the record directs the user to the data set.[43] The Council of European Social Science Data Archives and the GeoConnections Discovery Portal (IP2SF15) is an example of this type of portal.

Digital collections/portals can be housed in a single physical location (Statistics Canada – IP2SF18), and they may be virtual (Earth Systems GRID – IP2SF36), housed in a set of physical locations and linked electronically to create a single, coherent collection (Global Change Master Directory – IP2SF32, International Comprehensive Ocean Atmospheric Dataset – IP2SF25). The distinction between centralized, distributed, or unified portals may have funding, policy, and preservation implications. Data collections may also differ because of the unique policies, goals, and structure of the funding agencies. Collections created and maintained by government data centres, such as the NASA Space Mission Data (CS08), data federations like

40 NRC, *Spatial Information Resources*, pp. 65–66.
41 Ibid.
42 Ibid.
43 Ibid.

the Mammal Networked Information System, and specific research projects, such as the MOST Satellite Mission[44] (CS26), each pose unique challenges for policy-makers. Three functional data collections/portal categories are: 1) research data collections; 2) resource- or community-data collections; 3) reference-data collections. These are not rigid categories, particularly since some research-data collections are also resource collections. These collections are also indicative of the long-standing practice of collaboration in the sciences.[45]

Research Data Collections/Portals

Research data collections[46] or portals contain the results of one or more focused research projects and data that are subject to limited processing. Data types are specialized and may or may not conform or adhere to community standards, metadata standards, and content-access policies. Data collections vary in size but are intended to serve a specific scientific group, often limited to immediate participants. These collections are supported by relatively small budgets, often through research grants funding a specific project, and therefore do not have preservation as a priority. The Microvariability & Oscillations of Stars (MOST) Satellite Mission on-line database Case Study (CS26) falls into this category, as do a number of the portals in the GS10 study, including:

- IP2SF5 – IU (Indiana University) Bio Archive
- IP2SF6 – Computational Chemistry Archives/Computational Chemistry List
- IP2SF7 – The FMRI Data Center (fMRIDC) [Functional MRI]
- IP2SF8 – NIST (National Institute of Standards and Technology) StRD Statistical Reference Data Sets (Dataset Archives)
- IP2SF19 – National Virtual Observatory (NVO)
- IP2SF21 – Joint Center for Structural Genomics (JCSG)

Resource- or Community-Data Collections/Portals

Resource- or community-data collections[47] serve a single science, geomatics, or engineering community. These digital collections are often large enough to

44 Bart Ballaux, *CS26 Most Satellite Mission: Preservation of Space Telescope Data* (Vancouver, 2005).

45 American Institute of Physics (AIP), *AIP Study of Multi-Institutional Collaborations: Final Report. Highlights and Project Documentations* (Melville, 2001), http://www.aip.org/history/publications.html (accessed 17 August 2007).

46 NSF, *Report of the National Science Board*, p. 20.

47 Ibid.

establish community-level standards, either by selecting from among pre-existing standards or by bringing the community together to develop new standards where they are absent or inadequate. The CanCore Learnware metadata standard[48] is an example of this type of community standard. The budgets for resource or community data collections are moderate and often supported by a government agency. Preservation is contingent on departmental or agency priorities and budgets. Five of the GS10 Portals that fit this description include:

- IP2SF14 – Canadian Institute for Health Information (CIHI)
- IP2SF17 – National Institutes of Health (NIH)
- IP2SF24 – Southern California Earthquake Center (SCEC)
- IP2SF26 – National Geophysical Data Center (NGDC – NOAA)
- IP2SF36 – Earth Systems Grid (ESG) portal

Reference-Data Collections/Portals

Reference data collections[49] are intended to serve large segments of the scientific, geomatics, and education community. These digital collections are broad in scope and serve diverse user communities, including scientists, students, policy makers, and educators from many disciplines, institutions, and geographical settings. Normally they have well-established and comprehensive standards, which often become either de jure or de facto standards, such as the Geomatics ISO 19115 Metadata standard and the Federal Geographic Data Committee Metadata standards. Budgets supporting reference collections/portals are often large and come from multiple sources in the form of direct, long-term support; the expectation is that these collections will be maintained indefinitely, but not necessarily archived. Examples from the GS10 study include:

- IP2SF15 – Canadian Geospatial Data Infrastructure (CGDI)
- IP2SF18 – Statistics Canada
- IP2SF32 – Global Change Master Directory – Global Change Data Center

Scientific Data Quality

To an archivist, an authentic record does not have to be an accurate record, and an inaccurate record can also be an authentic record by consistently reproducing the same error; therefore authenticity alone does not "automatically

48 CanCore, *CanCore Metadata Initiative,* http://www.cancore.ca/en/ (accessed 27 January 2007).

49 NSF, *Report of the National Science Board*, p. 21.

imply that the content of a record is reliable."[50] Scientists, on the other hand, give primacy to data quality, which includes authenticity, normally articulated as provenance or lineage. Data accuracy is critical and the data need to be reliable. Data quality is normally articulated in a data set's metadata; without metadata or data-quality parameters, a scientist will not use, trust, or rely on that data. Each scientific discipline differs in how it defines scientific data quality, as is demonstrated from the Case Studies and the Portal Survey results; however, most include some or most of the following data quality elements: positional accuracy; attribute and thematic accuracy; completeness; semantic accuracy; and temporal information, reliability, lineage, logical consistency, and objectivity.[51] These quality elements are normally captured in metadata, and geomatics researchers argue that digital data archivists must consider data quality if they are to acquire data from the sciences. Indeed, the data quality of a record may be an important factor in the decision of what scientific data to archive. Clearly it will be very difficult for archivists to make appraisal decisions about scientific data on their own and they may not necessarily be able to do so according to typical archival practices, as Ken Thibodeau points out:

> The relevant framework of appraising scientific data sets, thus, is not defined by the business activities or the need for corporate memory of the sponsoring agency, but by the research community. Seeking the input of scientists in the appraisal of the data recognizes that the roles and the actions of academic researchers are at least as important as the functions of the agency that funded the research or launched the satellite.[52]

Data quality will be one of the important elements that will need to be factored into the archival appraisal process and will require the assistance of data creators and scientists themselves. In the best of all worlds, data quality would have been included in metadata at the beginning of a data set's life cycle. Archivists have a role to play with funding agencies, scientific institutions, and scientists in ensuring that archival practices are part of the research process from the very beginning, as discussed in a number of studies on this topic.[53]

50 Pearce-Moses, http://www.archivists.org/glossary/ (accessed 23 August 2007).

51 Anders Ostman, *The Specifications and Evaluation of Spatial Data Quality*, Proceedings of the 18th ICA/ACI International Conference at Stockhom, 1997 and Stephen C. Guptill and Joel L Morrison, eds. *Elements of Data Quality* (Oxford, 1995).

52 Kenneth Thibodeau, "Preserving Scientific Data on Our Physical Universe," *IASSIST Quaterly* (Winter (1995), http://iassistdata.org/publications/iq/iq19/iqvol194thibodeau.pdf (accessed 1 August 2007), p. 26.

53 See AIP, *Study of Multi-Institutional Collaborations* and NRC, *Preserving Scientific Data on Our Physical Universe*.

Wars, analysis and predictions of calamities, vacationing, real-estate transactions, medical research, exploration, etc., rely on accurate quality data. For many centuries, people have been willing to pay high prices for high quality data. Think of spies, planners, construction engineers, and especially those involved in the medical and military sciences who want more exacting data quality.[54] Currently, in the field of geomatics the

> ... cartographer and the data provider may not know each other, there are now many data producers and the user must choose among them and datasets look deceivingly more accurate, while all data sets include inherent inaccuracies that need to be accounted for by the user. It is now a professional obligation of cartographers to include knowledge about the quality of a data file used.[55]

In essence, quality can be associated with uncertainty and error.[56] Data accessed from portals is a classic representation of this same situation; the users may not know the scientist who produced the data, but they will decide whether or not a particular data set is fit for their use by examining the data quality elements in the metadata, while also subjectively assessing the degree of trust they have in the organization hosting the data portal.

The problem of preserving authentic and reliable digital data and records for the near and longer terms is not unique to the sciences. It faces everyone who now or in the future will require research data, legal documents, and administrative records to conduct their business, because more and more material is being created only in a digital form and will be communicated, stored, and accessed only in digital systems. Individual users, as well as governments, business, and the courts will need to have data and records that have been preserved as authentic information objects. The generality of the problem points to the widespread need for digital preservation strategies and procedures that are purposely designed to assure people that the data and records they rely upon are what they purport to be, and are free from tampering or corruption. Statistics Canada (IP2SF18), for example, states that "the confidence of clients in the quality of that information is critical to the Agency's reputation as an independent, objective source of trustworthy information."[57] Data and records in the sciences need to be authentic too; that is to

54 Joel L. Morrison, "Spatial Data Quality," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 1–12.

55 Ibid., p. 2.

56 Michael F. Goodchild, "Attribute Accuracy," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 59–79.

57 Statistics Canada, *Statistics Canada's Quality Assurance Framework* (Ottawa, 2002), http://www.statcan.ca/bsolc/english/bsolc?catno=12-586-X&CHROPG=1 (accessed 5 February 2007), p. 1.

say that their original identity and integrity have not been compromised, either on purpose or through inadvertence. In addition, in many of the sciences there is a further requirement that the individual datum and aggregated data and data sets be accurate, and that these data be maintained accurately over the long term.

The OMB *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies* document is insightful. Quality is considered to be "an encompassing term comprising utility, objectivity, and integrity."[58] As the guidelines state:

- *Utility* refers to the usefulness of the information to the intended users.
- *Objectivity* focuses on whether the disseminated information is being presented in an accurate, clear, complete, and unbiased manner, and as a matter of substance, is accurate, reliable, and unbiased.
- *Integrity* refers to security – the protection of information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.

Utility is closely related to the concept of fit-for-use; *objectivity* includes all of the elements of scientific-data quality, while *integrity* in this case is analogous to authenticity in archival science.

Further, these guidelines state that

> ... agencies shall develop a process for reviewing the quality (including the objectivity, utility, and integrity) of information before it is disseminated. Agencies shall treat information quality as integral to every step of an agency's development of information, including creation, collection, maintenance, and dissemination. This process shall enable the agency to substantiate the quality of the information it has disseminated through documentation or other means appropriate to the information" [and] agencies shall adopt specific standards of quality that are appropriate for the various categories of information they disseminate."[59]

Certainly, archivists mandated to preserve these data for the long term will also have to include data quality among the values they assess in the course of their appraisal processes. The National Geophysical Data Center (IP2SF26) portal makes explicit reference to the OMB data quality guidelines in its data management strategy as does the National Institute of Health (IP2SF16) when referring to influential scientific data.

58 Office of Management and Budget (OMB), *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies: Federal Register*, vol. 67, no. 36 (Washington, DC, Friday, 22 February, 2002), www.whitehouse.gov/omb/fedreg/reproducible2.pdf (accessed 17 January 2007), p. 8453.

59 Ibid., p. 8459.

Authenticity

The effect of digital technology, with its associated and rapidly changing hardware and software platforms and environments, has been to upset traditional systems that controlled production, validation, and preservation of records and data. The control that was derived from secure custody has been eroded. This is particularly important for sciences that rely on post factum audit, on the one hand, to replicate experiments and conclusions, and on the other hand, to validate the claims of discovery that may have an impact on the general public. As many have observed, technology has eroded the concept of the original, which was, in the past, tied to the notion of the first complete instantiation of recorded data that was effectively communicated to others across space or time.

The meanings of "authenticity" are relative to the concept of authentic that is held by different disciplines. Authenticity may mean different things to an artist, a lawyer, a musician, or a scientist. To be authentic, something must be what it claims to be; it must not be something other than what it claims, either as a result of a mistake, or misrepresentation. That which is authentic cannot be drawn up or be manufactured as a similacrum: the thing cannot be forged and be willfully misrepresented. Authenticity is tied to a person as the author or the creator, and to the processes of creating, or accumulating and acting upon: both person and process must be present and available for assessment by the reader, listener, or user.

The Society of American Archivists (SAA) Glossary defines authenticity as: "the quality of being genuine, not a counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context."[60] Authentic means "perceived of as genuine, rather than as counterfeit or specious; bona fide."[61] The InterPARES 1 Glossary defines authenticity as "the quality of being authentic, or entitled to acceptance. As being authoritative or duly authorized, as being what it professes in origin or authorship, as being genuine."[62] The IP2 Glossary adds that authenticity is "the trustworthiness of a record as a record; i.e., the quality of a record that is what it purports to be and that is free from tampering or corruption. [Archives]."[63] Pearce-Moses adds that "authen-

60 Pearce-Moses.

61 Ibid.

62 InterPARES 1, The *InterPARESGlossary* (December 2001), http://interpares.org/display_file.cfm?doc=ip1_glossary.pdf (accessed 6 September 2007).

63 InterPARES 2, *The InterPARES 2 Project Glossary* (Vancouver, 2007), http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=275089&CFTOKEN=92905534 (accessed 23 August 2007).

ticity is closely associated with the creator (or creators) of a record. First and foremost, an authentic record must have been created by the individual represented as the creator." Finally, the InterPARES 1 Project provides a set of Benchmark Requirements to attest to the authenticity of a record.[64]

When referring to these concepts, scientists do not normally use the term authenticity, but instead use the terms lineage,[65] data provenance or data integrity.[66] In the thirty-two portals surveyed, the term "authentication" is used, and many of the qualities of authenticity are discussed, while the term authenticity does not appear.

Authenticity is closely tied to the concept of trustworthiness. An object that is believed and proven to be authentic is considered to be trustworthy. Trust is the quality that underpins social relations, and business and juridically countenanced transactions. Trust is often built or erected on the guarantees that data or records are authentic, reliable, and accurate. These qualities, among other features, suggest that a thing or person is trustworthy.

The Benchmark Requirements for Authenticity discussed in the InterPARES 1 report are essentially the same elements that are found in good scientific data metadata (see the ISO 19115 Elements described in the InterPARES 2 MADRAS Registry[67]). The term authentication is often discussed in the sciences in the same way that a computer scientist would use it, namely referring to data kept secure by a process of passwords or by another security system, and have not been modified in or during transfer.

The responses to InterPARES 2 Case Studies questions regarding authenticity[68] revealed that the prevalent concept of reliability, discussed later in the paper, is closely tied to reproducibility and accuracy. In many cases, reliability is associated with faith in the technological systems in place or security measures related to accessing the system. Trust in data sources are important in both the Cybercartographic Atlas of Antarctica (CS06) and the Archeology

64 InterPARES 1, Authenticity Task Force, R*equirements for Assessing and Maintaining the Authenticity of Electronic Records* (Vancouver, 2002), http://interpares.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf (accessed 6 September 2007).

65 William Underwood, *CS08 Mars Global Surveyor Data Records in the Planetary Data System: A Case Study* (Vancouver, 2005).

66 OMB, p. 8459.

67 InterPARES 2 USA*, Metadata and Archival Description Registry and Analysis System (MADRAS) beta version* (Los Angeles, 2005).

68 See "Table 1: InterPARES 2 Selected Case Study Authenticity Responses" in the supplementary file attached to the electronic version of this article in *e-Archivaria*, or on the InterPARES 2 website at http://www.interpares.org/ip2/ip2_dissemination.cfm?proj=ip2&cat=pu-atcl-r .

records study (CS14),[69] while technological integrity, validity procedures and system checks are implemented in the Mars Global Surveyor Data Study (CS08), the Computerization of Alsace-Moselle's Land Registry (CS18), and the Most Satellite Mission project (CS26). Some of the studies control access via the appointment of responsible agents (VanMap [CS24] and the Alsace-Moselle Land Registry), or specialists (Cybercartographic Atlas). Peer review of data is used as a method for the Mars Global Surveyor Data study. In the Engineering Objects Study (CS19) the creators do not believe their records to be authentic as they have no assurance system that they are. Based on their responses, it would seem that the case-study respondents do not think of authenticity in the same way that archivists do, as there seems to be more an emphasis on measures to ensure data quality. Most creators are trying to ensure that they have good quality data and not necessarily authentic data in an archival sense. The Land Registry may be the exception, as the system is specifically designed to ensure that each registration is authenticated in the system.

The GS10 portals that were surveyed provided a rich array of information on the topic of authenticity in the sciences. Ensuring the data are of good quality and can be trusted is critical to these data portals or users would not rely on them. For many of the portals, the process of data control begins when the data are ingested into the system, made accessible via Web server sharing protocols, or described into a metadata description form. Some portals only ingest data that are derived from peer review journals (National Virtual Observatory – IP2SF19). Others only allow certain groups, organizations, or individual researchers to contribute, such a modelers (Earth Systems Grid – IP2SF36), consortium members (Joint Center for Structural Genomics – IP2SF21, The FMRI Data Center – IP2SF7), members of approved research projects (IP2SF2 – NASA Life Sciences Archive – IP2SF2), designated scientists (OBIS-SEAMAP Ocean Biogeographic Information System – Spatial Ecological Analysis of Megavertebrate Populations – IP2SF13) and approved government programs (National Snow and Ice Data Center, NASA – IP2SF28). Some acquire data from purchased sources (IP2SF14 – Canadian Institute for Health Information). Finally, others restrict data only on the basis of whether they fit the mandate of the community that the portal serves.

Once the data are in a particular portal, there are a wide variety of security measures in place to ensure they are not tampered with. Many portals include user authentication and/or registration mechanisms (USGS Data Portals –

69 Erin O'Meara, Richard Pearce-Moses, and Randy Preston, *CS14 Archaeological Records in a Geographical Information System: Research in the American Southwest* (Vancouver, 2004).

GEO-DATA Explorer – IP2SF37; Community Data Portal at NCAR – IP2SF35); others only have trained personnel working the system (The National Cancer Registry – IP2SF16), Public Key Infrastructure (PKI) and Department of Environment (DOE) GRID Certification Authority access (Earth Systems Grid – IP2SF36), and authorized users (Long Term Ecological Research – IP2SF23). Data users have metadata that include lineage information to help them determine if the data are fit for use. Before the data are made available to the public or to specific research communities, some validation processes are in place to attest to their authenticity and quality (NASA Life Sciences Archive – IP2SF2; Southern California Earthquake Center – IP2SF24). However, there did not seem to be any mechanisms in place that discussed how users can assess that the data sets that they have downloaded or received are authentic beyond metadata and file headers.

Presumption of Authenticity

As we observed earlier, the creators of Science Case Studies, the VanMap Case Study, as well as of most of the Data Portals surveyed, do not use the term authenticity. Where metadata exist, along with secure access to the data, many of the InterPARES 1 benchmark authenticity requirements are met. However, where there are no metadata, or in the case where a user wishes to assess if a data set has been tampered with during transfer, the concept of *presumption of authenticity* is introduced. Defined as "an inference as to the fact of a record's authenticity that is drawn from known facts about the manner in which that record has been created and maintained,"[70] this presumption is of value for the case studies. For example, the SAA definition notes that "the authenticity of records and documents is usually presumed, rather than requiring affirmation. Federal rules of evidence stipulate that to be presumed authentic, records and documents must be created in the 'regular practice' of business and that there be no overt reason to suspect the trustworthiness of the record (Uniform Rules of Evidence, as approved July 1999)."[71] This being the case, the context, practices, associated documentation, validation processes and authentication, and access measures would suggest that Case Studies discussed in this paper and the data in the GS10 portals are presumed authentic from an archival perspective.

70 InterPARES 2, *Terminology and Glossary* (Vancouver, 2007), http://interpares.org/ip2/ip2_terminology_db.cfm (accessed 17 January 2007).

71 Pearce-Moses.

The presumption that a valuable and authentic archive of data will emerge over time rests on a foundation of trust built between the original provider, his or her successors, including archivists and archival agencies, and the end user. That foundation is anchored in two pillars: 1) an explicit description of the sources of the data, and of the changes and processes that the data have undergone over time, so that any user is able to come to a decision about whether the data fit their proposed use; and 2) the continuing authority that the portal maintains as a viable community of practice and data. This is largely a result of a robust mandate, a stable sponsorship that provides an assured source of funding, open and accessible policies governing access to the data, and a declared or understood commitment to the public good. The portals previously mentioned under authenticity, and many others, more or less fit this bill.[72]

Lineage

Lineage is information that describes the source of the observations, data collection and compilation methodologies, conversions, transformations, analyses, and derivations to which the data have been subjected. It also provides the assumptions and the criteria applied at any stage of its life, as well as any biases. In fact, lineage is normally the first part of a quality statement, since most other data quality elements are affected by lineage.

Data producers have documented procedures and quality requirements that they have to meet, and lineage is a kind of audit trail to attest to the fact that the producers have met those requirements. Lineage provides a data set with its pedigree and allows the user to decide on its fitness for use; it can also be found in a data set's associated publications, reports, and technical notes. The "ultimate purpose of lineage is to preserve for future generations the valuable historical data resource. The key to our understanding of the Earth system may lie in the data collected by past generations."[73]

Accuracy

If accuracy can be considered to represent distance from the truth, then the truth should be known. But the truth cannot be known; it is instead accepted that the true position that could be obtained using the best available surveying techniques, personnel, uptodateness, etc.[74]

72 Tracey P. Lauriault and Barbara L. Craig, "Do Data Access Portals, Repositories, and Catalogues Preserve or Archive Geospatial and Science Data?" *Geotech 2006* (Ottawa, 2006).

73 Clarke and Clark, p. 15.

74 Jane Drummond, "Positional Accuracy," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), p. 34.

The measure of accuracy, error, or distance from the truth is a critical data quality element for scientific data. There are numerous sources of error in any given data set, and data users are well aware of these; therefore, they count on metadata descriptions that include data lineage and accuracy measures to enable them to make the decision to trust and use those data. Alternatively, scientists will refer to peer reviewed papers that discuss the scientific model applied and the research methodology directing data collection. There is no accuracy measure that can be used in all situations. Each scientific community and each specific data set includes its own accuracy parameters and particularities. For scientists, "accuracy can be defined as the difference between a measurement, or attribute, and some comparable measurement known to be of higher accuracy."[75] This argument, however, is circular and relativist, therefore data can "never be more accurate than the most accurate source."[76] Accuracy is the most pervasive and common metadata element; "to a purist, no number has meaning unless it is accompanied by an estimate of uncertainty … [and] at a minimum, the metadata should include general comments on the maximum expected errors, even if a quantitative measure such as standard deviation cannot be given."[77] In the archival community, accuracy is less specific than in the sciences, and is defined as "the degree of precision to which something is correct, truthful, and free of error or distortion, whether by omission or commission,"[78] or "the degree to which data, information, documents or records are precise, correct, truthful, free of error or distortion, or pertinent to the matter [Archives]."[79] In the sciences, errors are a given, therefore, a measure of the margin of that error is imperative.

As an example, in the field of geomatics "positional accuracy represents the nearness of those values to the entity's 'true' position in that system,"[80] position being defined by a coordinate system or a Projection Grid Coordinate System. On a map, land areas and points will adhere to different systems, such as a geodetic coordinate systems to store positional entities (e.g., wiggle of a river, a telephone pole). A number of transformations can occur from the point of collection, from one grid system to another and from one Geographic Information System (GIS) to another, and for a variety of different visualizations; this, of course, introduces errors. Users want data that are very near the positional truth, and in some case will accept a certain distance from the truth depending on the use of the data.

75 Goodchild, p. 66.
76 Ibid.
77 NRC, *Preserving Scientific Data*, p. 37.
78 Pearce-Moses.
79 IP2, *Terminology and Glossary*.
80 Drummond, p. 32.

A data set's life cycle developing from acquisition to compilation and derivation comprises important areas of concern to accuracy.[81] *Acquisition* is most important. It is the point where the original observations are collected and where "fundamental assumptions, calibrations and corrections are made."[82] *Compilation* is the part where a database is created; it occurs when the facts are assembled into some sort of comprehensive arrangement or into a scientific data set, and it is a phase where many errors can be introduced. *Derivation* is the stage where data are being manipulated; the output of this process is a representation, interpolations, averaging, and any number of manipulative techniques that may change the form, format, or structure of the data. This may or may not be a reversible phase and is a diversion point from the original observations. For this reason, keeping the raw data as well as derived data is important.

Error

There are numerous causes of data errors in scientific records. During the *data collection* stage, errors may be related to the method of collection or the sensor collecting the data. During the *scientific parameter generating stage*, where observed values are transformed or data processed, these activities may introduce errors, including algorithmic errors, erroneous processing techniques, calibration issues, theoretical models, and boundary conditions. The *data conversion stage* can include data classification, contouring, and interpolation, to name a few. Since these are very subjective processes and subject to much interpretation, the possible biases need to be included in the data lineage. Errors at this stage depend on the equipment used for the conversion, operator policies, digitization policies, and the quality of the source material prior to conversion. In many cases there are numerous algorithms for the same type of transformation: these may yield different results and are often software dependent. Finally, the *production stage*, consisting of generating final products, such as maps, images, charts, and reports with data tables, may include cumulative sources of error resulting from any of the earlier stages.

A number of techniques can be used to assess the cumulative errors introduced into a data set; in the spatial sciences, a few of these techniques include variance propagation, direct measurement of errors, error modelling, error visualization, or other statistical methods. Accuracy is specified in uncertainty metrics, reliability diagrams, confidence measures, etc. Because these different methods to gather data are well-known to scientists, lineage becomes

81 Clarke and Clark, pp. 13–30.
82 Ibid., p. 18.

important; including methodology in lineage metadata alerts a data user to the level of error inherent in the data set under examination. On a map, for example, errors can affect where a line is drawn and a boundary exists or the size of an area. This is a seemingly innocuous detail, unless of course decisions are being made based on a line marking a route in an in-car navigation system, one is demarcating a country's border crossing with a fence, drilling for the construction of a tunnel, or logging in a protected bioreserve that was improperly demarcated.

Errors can be detected after a routine validation check; the Cambridge Crystallographic Data Centre (IP2SF4), for example, uses an encipher, Prequest, and other international CCDC software to test its data. If errors are detected, they are either resolved in-house or referred back to the original author(s) for clarification. The portal also has scientific editors, who provide text remarks concerning the resolution of errors and the structure, and record the nature of any crystallographic disorder.

Some of the data portals are very upfront with the imperfections of their data sets. The Ocean Biogeographic Information System – Spatial Ecological Analysis of Megavertebrate Populations (IP2SF13) makes explicit the most common errors:

- Observation points far inland;
- Observation points in an area that the species is not supposed to occur;
- Species with wrong taxonomy.

Users are also expected to recognize that the analysis and interpretation of data require background knowledge and expertise about marine biodiversity (including ecosystems and taxonomy). This expectation reinforces earlier observations that archivists will fare better at archiving specific types of scientific data if they collaborate with scientists and specialists in the field. Alternatively, archivists can trust that either the scientists or the body managing the portal will have already appraised the data in their portals, and they can instead work with portal managers and their related institutions to add specific archiving practices into the process.

Bias and Objectivity

Bias and objectivity are also key concepts in scientific enquiry, and it is useful to return to the US OMB Guidelines where objectivity "involves a focus on ensuring accurate, reliable, and unbiased information. In a scientific, financial, or statistical context, the original and supporting data shall be generated, and the analytic results shall be developed, using sound statistical and research methods."[83] The Guidelines distinguish between presentation and substance:

83 OMB, p. 8459.

- ''Objectivity'' includes whether disseminated information is being presented in an accurate, clear, complete, and unbiased manner. This involves whether the information is presented within a proper context. Sometimes, in disseminating certain types of information to the public, other information must also be disseminated in order to ensure an accurate, clear, complete, and unbiased presentation. Also, the agency needs to identify the sources of the disseminated information (to the extent possible, consistent with confidentiality protections) and, in a scientific, financial, or statistical context, the supporting data and models, so that the public can assess for itself whether there may be some reason to question the objectivity of the sources. Where appropriate, data should have full, accurate, transparent documentation, and error sources affecting data quality should be identified and disclosed to users.
- ''Integrity'' refers to the security of information—protection of the information from unauthorized access or revision, to ensure that the information is not compromised through corruption or falsification.[84]

Objectivity related to OMB's dissemination definition is an excellent high-level definition of what scientists consider to be essential to any data sets, namely lineage, accuracy statements, and accompanying documentation such as metadata. Integrity in this case is associated with both authentication and authenticity as defined by the archival community. In other words, integrity ensures that the data are what they purport to be, have not been modified in transfer, and a mechanism is in place to ensure that they meet the IP2 benchmarks for authenticity.

Reliability

In the sciences, the concept of reliability is closely associated with the concepts of reproducibility and accuracy. It can be related to the degree to which a forecast's or model's probabilities or results match the observed frequencies of an occurrence in the environment or consistently produce the same result. More generally, reliability is a quality that can be attributed to a person, as in a reliable person; to a device, such as a reliable machine; or to a system that is organized to accomplish certain ends, as in a reliable computer or records system. It is the individual assessor who determines what attributes are required before reliability can be reasonably inferred.

84 Ibid., pp. 8459–60.

The concept of reliability is similar for archivists. According to the SAA definition, reliability is "the quality of being dependable and worthy of trust. – 2. The quality of being consistent and undeviating. – 3. Diplomatics · Created by a competent authority, according to established processes, and being complete in all formal elements."[85] Reliability is considered to be one of the foundations of trustworthiness.[86] Trustworthiness thus has qualitative dimensions – reliability and authenticity – and quantitative dimensions – accuracy and completeness. If the record's integrity appears to be compromised in some way, or if its lineage is not clear and complete, knowledgeable users would have grounds for withholding trust.

In the sciences, there are a number of methods used to document reliability; in cartography, for example, a reliability diagram includes the authority that produced the map and the quality of the source material. Some rely on reliability diagrams that speak to the probability that a particular model, data, or experiment is accurate. Reliability measures are also closely associated with measures of error in a system or data set, which of course is in turn associated with a degree of accuracy previously discussed. Measures of reliability are statistically complex and designed to test the probability of forecasts or a model's outcome. The sections on errors attest to a data set's accuracy, and reporting on authenticity findings from the case studies shows computational methods to detect errors and to ensure reliability.

Other Concepts Associated with Data Quality

In the sciences, there are numerous other elements associated with data quality. A few particular to the spatial sciences are mentioned here for illustrative purposes only, as it is not possible within the scope of this paper to do them justice. One is *completeness*, which can be related to technical issues in algorithms or fundamental questions about the mental model and the scientific concepts used to represent a real-world phenomenon.[87] There are discrete measures of completeness, since this element is relative to a comparison object or an abstract model and is always context contingent. A data set that is incomplete needs to be described as such, and this enables the user to troubleshoot around what is missing and to modify the descriptions of its derivative products. Other elements such as *logical consistency,* address the structur-

85 Pearce-Moses.

86 Heather MacNeil, *Trusting Records: Legal, Historical, and Diplomatic Perspectives* (Dordrecht, 2000), p. xi.

87 Kurt Brassel, Felix Bucher, Eva-Maria Stephan, and Andrej Vckovski, "Completeness," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 81–108.

al integrity of a data set, and mathematical theories of metric, topology, and ordered sets are critical to a framework that is built on modelling the data and the relationships among objects.[88] *Semantic accuracy* in cartography refers to "the quality with which geographical objects are described in accordance with the selected model,"[89] (addressed later in more detail in the section on metadata and the discussion of formal ontologies); the element refers "to the pertinence of the meaning of the geographical object rather than the geometrical representation."[90] *Temporal information* is important for tracking the changes made to a data set, but also to ensure that when integrating disparate data sets or thematic time series they match and make sense. All of this relates to issues of lineage, positional accuracy, and attribute accuracy, etc.[91] Finally, *attribute accuracy* is a complex concept that relates to how a data entity is described and how that entity can accurately be represented or modelled.

Data Quality Disclaimers

Ironically, while most organizations aim to ensure their data are accurate, reliable, and authentic, the case studies and the data portals that we examined demonstrated that many of these same organizations will add disclaimers to absolve themselves of any responsibility for damages that may result from the use of their data. In the VanMap study (CS24), a special disclaimer is used in connection with utility data, stating that

> ... the City of Vancouver assumes no responsibility for the accuracy or completeness of the field information shown in VanMap. All work carried out is done wholly at the risk of the party undertaking the work who agrees, as a condition of such undertaking, to release the City of Vancouver from all liability. Location of underground utilities should always be confirmed by manual digging.[92]

While the Cybercartographic Atlas explicitly states that it is to be used solely for educational purposes, the Antarctic Digital Database (ADD) (IP2SF27) warns that its maps, when combined, may reflect some inconsistencies, particularly when older data sets are included. The National Geophysical Data Center (IP2SF26) and the World Data Center for Solar

88 Wolfgang Kainz, "Logical Consistency," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 109–37.

89 François Salgé, "Semantic Accuracy," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), p. 139.

90 Ibid.

91 Stephen C. Guptill, "Temporal Information," in *Elements of Data Quality*, eds. Stephen C. Guptill and Joel L. Morrison (Oxford, 1995), pp. 153–66.

92 Evelyn McLellan, *CS24 City of Vancouver Geographic Information System (VanMap)* (Vancouver, 2005).

Terrestrial Physics (IP2SF10) indicate that the government of the United States and its employees cannot be held accountable for any data quality warranties, and also ask that if errors are identified that they be notified. The FMRI Data Center (IP2SF7) absolves itself from liability in relation to data quality, while the IU (Indiana University) Bio Archive (IP2SF5) reminds users that data contain errors, and the British Atmospheric Data Centre (BADC) (IP2SF1) absolves itself from responsibility for data on the site and once downloaded onto the user's computer.

Data Quality in General Study 10 Portals

The preservation of accurate and authentic data over the long term is best seen as a complex problem that needs to be addressed on several levels. There are three groups of issues surrounding portals and data quality: those related to the portal's operation and its design, management, and long-term viability; those related to the accuracy of the individual datum and data sets; and those related to the relationship between the portal, its data and services, and the individual or corporate user – essentially those issues that emerge from a history of interaction that builds trust and comfort with the user.

The issues that are related to the portal itself are those that are linked to maintaining an authentic memory, especially of the sources of the data, their management or changes over time, and their connections to contributors or sources. Building sites and services that continue to be what they purport to be, and whose changes and transitions over time are visible and knowable to a user build conditions of trust. The InterPARES 1 project developed benchmarks that could be used by portals to ensure that their data continue to be authentic over time.

The portals investigated for General Study 10 represent the heterogeneity of scientific research. Data quality, as we have seen, has many dimensions: authenticity, as it is customarily viewed within archives, is only one of these dimensions.

Amongst the portals examined, there are several different examples of how data-quality issues are addressed. The British Atmospheric Data Centre (BADC) (IP2SF1) follows the terms and conditions of its host organization, the Natural Environment Research Council, and produced the document *Terms and Conditions for data and information provided by the NERC/BADC.*[93] Combined with disclaimers on use discussed below, the document includes discussions of data limitations:

93 Natural Environment Research Council (NERC), *Terms and Conditions for data and information provided by the NERC/BADC* (Oxfordshire, 2007), http://badc.nerc.ac.uk/conditions/badc_anon.html (accessed 27 January 2007).

- Issues related to scientific models that evolve over time, which affect how data are collected;
- Data errors introduced during transcription and transformation;
- Details related to scale differences;
- Third-party data which may not have been reviewed;
- Some data sets collected to serve particular purposes and which may be incomplete for other uses.

The National Institute of Standards and Technology Statistical Reference Data Sets' (IP2SF8) entire mandate is "to improve the accuracy of statistical software by providing reference data sets with certified computational results that enable the objective evaluation of statistical software." It stores data that can be used by statistical software developers to test the robustness of the algorithms in their software. Each data set is accompanied by accuracy parameters for a particular software, and a number of precision methods are in place for different statistical formulas, such as analysis of variance, linear regression analysis, non linear regression, and so on. Precisely because the data sets in this portal are used for testing and evaluation of statistical software and computational accuracy, much attention has been paid to data quality, in particular accuracy. This is the case for the data, the certified values, and the algorithms used to manipulate the data and the software packages. The Ocean Biogeographic Information System – Spatial Ecological Analysis of Megavertebrate Populations portal (IP2SF13) includes a list of major data gaps that mostly address completeness of its maps: the deep sea is the least surveyed part of the planet; coastal areas have not been adequately sampled for the distribution of wildlife; northern oceans are more sampled than those in the south; many marine species are not named; there have been naming changes overtime; some species data have not been published; many have not been entered into databases; and many databases are not connected to OBIS – SEAMAP. Completeness is also an issue for the National Cancer Registry (IP2SF16) since they determined that their data only cover the primary cause of death, and some cancers may not appear. These qualifications allow the users to make cautious and informed assumptions on the data they are using. The Canadian Institute for Health Information (IP2SF14) includes in its data-quality framework five parameters: accuracy; comparability; timeliness; usability; and relevance. The National Virtual Observatory (IP2SF19) provides a unique method to assess data quality:

> ... [in the] VO architecture, there is **nobody deciding what is good data and what is bad data**, (although individual registries may impose such criteria if they wish). Instead, we expect that good data will rise to prominence organically, as it does on the

World Wide Web. We note that while the Web has no publishing restrictions, it is still an enormously useful resource; and we hope the same paradigm will make the VO registries useful.[94]

While the Long Term Ecological Research (IP2SF23) portal suggests that responsibility lies with data providers, it contains a number of Quality Assurance (QA) Controls, such as General Guidelines for QA, parameter-specific guidelines, and parameter-specific, default-threshold values and checks, which are also included in the metadata.

Some portals will provide assurances for their own data, but not for data from third-party organizations (World Data Center – IP2SF10, OBIS-SEAMAP – IP2SF13), or indicate that the data quality rests with those who provide or submit their data to the portal (GCDI – IP2SF15, Long Term Ecological Research – IP2SF23, Global Change Master Directory – IP2F32).

Data Quality in IP2 – Selected Case Studies

We now turn to data quality, particularly accuracy, within the IP2 Case Studies.[95] Again, our research shows a varied approach to addressing the issue of data quality. The Cybercartographic Atlas (CS06) relies on the professional practices and authority of the institutions from which data are derived, and adheres to cartographic professional practices to choose the right level of data accuracy and to select cartographers for the right representation, a process that is very much reliant on metadata and professional practices. The Mars Global Surveyor Data Records in the Planetary Data System (CS08) includes data processing plans, manuals, specifications, and workbooks to guide processing, transferring, and data preparation. Further, the data are peer reviewed for accuracy and reliability, and are validated through a system that also conducts checksums. Accuracy in the Coalescent Communities GIS (CS14) is more subjective and specific to the person who decides which data sets will be used; data sourcing is less formalized and there is a range of error acceptable to the profession. For the Alsace-Moselle Land Registry (CS18), a rigorous system including data verification, validation processes, PKI signatures and cross referencing, along with a well-designed architecture, ensures that the registries are accurate and authenticated within a legal-evidential framework

94 International Virtual Observatory Alliance, *Virtual Observatory Architecture Overview Version 1.0* (2004), http://www.ivoa.net/Documents/Notes/IVOArch/IVOArch-20040615.html (accessed 23 August 2007), emphasis in original.

95 See "Table 2: Accuracy Statements from the InterPARES 2 Case Studies in the Sciences and Geomatics" in the supplementary file attached to the electronic version of this article in *e-Archivaria*, or on the InterPARES 2 website at http://www.interpares.org/ip2/ip2_dissemination.cfm?proj=ip2&cat-pu-actl.r .

of property ownership. In Authenticating Engineering Objects (CS19), accuracy is in the hands of designers and their adherence to design modelling standards and CAD system geometry checks. For VanMap (CS24) there is a distributed system for establishing data accuracy based on from whom and where the data come. There is much reliance on professional practice, the authority of organizations from which external data sources are derived, and, before work is conducted that relies on the data, the data are ground truthed.[96] The Most Satellite Mission (CS26) has put in place a technical validation process with a series of checksums and scientists verify the data. Some data are processed through a model; any inaccuracies are addressed at that stage, and the data are potentially reprocessed.

Observations derived from these few case studies suggest that accuracy is associated with the risk of having inaccurate data: the more legal requirements there are, the more rigorous are the quality checks, as in the case of the Alsace-Moselle Land Registry (CS18). Also, the more automated the process is, the more technical the checksums are and the more reliant the creators are on the technical systems in place, the less reliant they are on human checks: this is the case with the NASA Mars Surveyor Data, the Engineering Drawing study, and the MOST satellite data. Professional practice, however, is very important in the Cybercartographic Atlas of Antarctica and the VanMap studies, as is a reliance on the trust associated with the integrity and authority of external data providers. The Atlas, however, relies heavily on good metadata provided by external data sources and on the metadata related to the atlas modules themselves. Finally, the Archeology study included the widest margin of error in its practices and the most subjective quality checks.

Metadata

To make data useable it is necessary to preserve adequate documentation relating to the content, structure, context, and source (e.g., experimental parameters and environmental conditions) of the data collection – collectively called metadata. Ideally, the metadata are a record of everything that might be of interest to another researcher. For computational data, for instance, preservation of data models and specific software is as important as the preservation of data they generate. Similarly, for observational and laboratory data, hardware and

96 Ground truthing is a process by which a feature on a map or a satellite image is compared to what is there in reality (at the present time). It can be used to verify the accuracy of a classified image, or to calibrate the pixels of satellite images to real features and materials on the ground. This is done to minimize errors in the classification, such as errors of commission and omission, or to validate a feature labelled in an image.

instrument specifications and other contextual information are critical. Metadata is crucial to assuring that the data element is useful in the future. The use of metadata and their accuracy have increased over the past several decades.[97]

As discussed throughout this paper, metadata are essential for the dissemination of scientific data whereby "a data set without metadata, or with metadata that do not support effective access and assessment of data lineage and quality, has little long-term use."[98] Authenticity in the sciences is linked to a clear lineage recorded in the accumulating metadata surrounding data. Both data and their cumulative and related metadata must be present, clear, unambiguous, and uncompromised. Lineage information supports assessments of the probability of error, either in the data, or in its collection, compilation, aggregation, or derivation.

Data portal discovery services rely on metadata descriptions. Metadata is like a form of truth in labelling and it is considered "axiomatic that a database has limited utility unless the auxiliary information required to understand and use it correctly – the metadata – is included in the record."[99] The data quality elements discussed throughout this text are captured in metadata. Scientists will not trust a data set that does not come with a description, and one cannot determine if a data set is fit for a particular application without metadata. The major uses of metadata include "(1) managing and maintaining an organization's investment in data, (2) providing information to data catalogs and clearinghouses, (3) providing information to aid data transfer and use, and (4) providing information on the data's history or lineage."[100] Metadata are also a means of attesting to and assessing a data set's authenticity. In the absence of metadata, it is possible to gain some understanding of a scientific data set if there are associated peer review papers and reports that describe them; however, this would be a more laborious process.

Metadata schemas and standards in the sciences, particularly geographic information science, are well-developed, and are essential for data discovery and fit-for-use decisions, however, many metadata standards remain housed in communities of practice and domain specific classification systems and data structures.[101] This is very apparent in the selected case studies and the portals we examined in General Study 10. We further observed that existing archiving metadata could be expanded to include metadata standards from other

97 NSF, *Report of the National Science Board*, p. 20.
98 NRC, *Preserving Scientific Data*, p. 36.
99 Ibid., p. 31.
100 Ibid., p. 62.
101 A. Gupta, B. Ludascher, M.E. Martone, "Registering Scientific Information Sources for Semantic Mediation," *Lecture Notes in Computer Science* no. 2503 (2002), pp.182–98.

disciplines; the IP2 Description Team is exploring this issue by analyzing the ISO 19115 *An International Metadata Standard for Geographic Information*. Therefore, information models (i.e., ontologies) derived from formal and informal methods could be used to assess the feasibility of maintaining knowledge of data-use context over time.

Formal Ontology

As stated, an essential part of creating effective metadata is the establishment of the context within which the data were collected or generated. Although it is important to document instrumentation parameters and methods used in data construction, establishing semantic context is key to understanding the meaning of scientific data. In this instance, semantics refers to how – typically through the use of language – computer-based representations stored in information systems are related to entities and concepts in the real world.

At the time of data collection, the semantic quality of the data may seem intuitive and well understood by all. Data producers, users, and stewards will likely understand the fundamentals and nuances of the lexicon or jargon surrounding the data. However, meanings tacitly understood and accepted at the time of creation may become obscure or change. Moreover, as most data are generated within specific domains and communities of practice, the semantic aspects of a data set may be quite specific and difficult for outsiders – such as archivists – to comprehend. Thus, the problem of semantic heterogeneity is introduced at the time of data creation or as the data age. Semantic heterogeneity is produced when a different symbol is used to convey similar meaning (e.g., "trunk" vs. "boot" when referring to the storage compartment in an automobile), or, when the same symbol is used to convey similar meaning (e.g., tank used to refer to a military vehicle or a liquid-storage container).

While semantics and semiotics have been integral to the disciplines of philosophy, linguistics, library science, and other disciplines, formal modelling of semantics within a computer environment has emerged more recently in the domain sometimes referred to as "knowledge engineering" or "knowledge representation." An increasingly important tool used for knowledge representation is the "formal ontology." Traditionally, "ontology" refers to the branch of philosophy that studies the nature of being, reality, and substance. In simplest terms, a formal ontology can be defined as a specification of a conceptualization.[102] More specifically, a formal ontology is a controlled vocabulary that describes objects in a domain and the relations

102 Tom R. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, vol. 5, no. 2 (1993) pp. 199–220.

between them using formal constructs (e.g., first order predicate calculus) stored and processed within a machine-based, digital computation environment.

Controlled vocabularies are typically an important element of a metadata schema. Thus, the formal ontology is a particularly relevant extension to simple controlled vocabularies. Both are used to capture and constrain intended meaning of a domain vocabulary; in the case of a formal ontology, however, the vocabulary is not simply presented and defined (as may be the case with a controlled vocabulary), but very explicitly explained using the combination of definitions, structural information (e.g., position within a hierarchy), and logical descriptions. The structural information and logical descriptions are used to infer meaning through formal reasoning.

Formal ontologies may have an important role in the long-term preservation of scientific data for the following reasons:

a) A formal ontology is explicit and analytic. If rigorously constructed using input from relevant stakeholders, the resulting detail and precision can provide considerably more information than a glossary or taxonomy.
b) Through the use of reasoning tools, the formal specifications inherent to an ontology can be inferential. If A=B and B<>C, then A<>C. Thus, an ontology as presented may provide partial context, while the inferences that can be made from the ontology can extend this contextual information.[103]
c) The inferential capabilities alluded to in (b) can be used to mediate semantic heterogeneity, thus supporting efforts in semantic translation.
d) Formal ontologies are typically expressed using established, well-defined structures (i.e., predicate logic) and are typically stored using simple technical and syntactical devices (XML using ASCII or Unicode Transformation Format UTF-8).
e) Formal ontologies are increasingly being developed and used by scientific communities.

In relation to item (d), for example, the participants in Cybercartographic Atlas of Antartica Case Study (CS06) have been involved in a number of initiatives within the polar-science community that are currently examining, developing, or using formal ontologies. These include: the extension of the Scientific Committee on Antarctic Research (SCAR) Feature Catalogue to a

103 It is recognized that formally modelling the vocabularies related to some scientific data sets may be difficult or impossible. Lack of consensus around domain semantics may present difficulties when attempting to establish a logic model.

formal ontology[104]; the investigation of formal ontologies for use by the Joint Committee on Antarctic Data Management,[105] the development of ontologies for the marine community (including the polar oceans), the Marine Metadata Initiative[106]; and the formation of the International Polar Year Knowledge Organization Group.[107]

The use and implications of formal ontologies and emerging forms of knowledge representation for scientific data preservation are areas of ongoing research within projects emerging from the Cybercartographic Atlas; they are being explored as a curation method by the UK Digital Curation Centre and are being tested by the Council for the Central Laboratory of the Research Councils.[108]

The selected case studies demonstrate that each scientific and geomatics project adheres to its own standards, some of which are developed by a community of practice, as in the case of the Cybercartographic Atlas (CS06), which adheres to ISO 19115, DIF, and FGDC metadata standards.[109] The Mars Surveyor NASA Data Study (CS08) adheres to a NASA institutional and data type specific Planetary Science Metadata standard, while Authenticating Engineering Objects (CS19) adheres to strict corporate and vendor standards. The MOST Satellite Mission Study (CS26) developed its own very basic standard to meet the needs of its project. VanMap (CS24), however, does not have a clear standard, and the Archeology study (CS14) simply refers to the source of the data ingested into the GIS. The Land Registry study (CS18) indicates there are no metadata, which may be related to the secure and encrypted access protocols and the architecture of the system. However, it is assumed that some basic catalogue type of metadata

104 Peter L. Pulsifer and A.P.R. Cooper, "A Geographic Grammar for Antarctica: Features, Semantics and Ontology," *SCAR Open Science Conference (*Hobart, 2006).

105 Joint Committee on Antarctic Data Management, *Report of the Tenth Joint Committee on Antarctic Data Management Meeting (JCADM-10)* (Hobart, 2006), www.jcadm.scar.org/fileadmin/documents/jcadm10/JCADM10Report.doc (accessed 17 January 2007).

106 NSF, *Marine Metadata Initiative* (Arlington, 2007), http://marinemetadata.org/ (accessed 27 January 2007).

107 International Polar Year Knowledge Organization Group, *International Polar Year Data Management Workshop, 3–4 March 2006* (Boulder, 2006).

108 For a more detailed treatment of the topic, the reader is directed to the relevant literature. See: N. Guarino, "Formal Ontology in Information Systems," *Proceedings of the First International Conference (FOIS'98)*, (Paper read at FOIS'98 Trento, June 1998); John F. Sowa, *Ontology* (2003), http://www.jfsowa.com/ontology, (accessed 21 January 2007); Pragya Argawal, "Ontological Considerations in GIScience," *Journal of Geographical Information Science*, vol. 19, no. 5 (2005), pp. 501–36.

109 Federal Geographic Data Committee, *Geospatial Metadata Standards* (2007), http://www.fgdc.gov/metadata/geospatial-metadata-standards#whatstandard (accessed 27 January 2007).

elements exist in order to find and retrieve registrations in the system. This small sample once again demonstrates the specificities inherent in the sciences.

Most but not all of the data portals we examined include metadata. Some are very minimalist and include only header files (Cambridge Crystallographic Data Centre – IP2SF4); others refer to associated peer review articles; some were designed specifically for a particular data set, while others adhere to the metadata standards of their discipline (Canadian Geospatial Data Infrastructure – IP2SF15), access portal, or institutions (World Data Center for Solar Terrestrial Physics – IP2SF10).

Scientific Records

The archival discipline and profession have a long and distinguished history and well-established traditions, theories, methods and practices, which are central to InterPARES 2. The project has, however, made explicit efforts to involve academics, professionals, and practitioners from a variety of other disciplines. Not surprisingly, each discipline has its own traditions, which may conflict with those of archival science. The interdisciplinary research process is a challenging, mutual learning process and often a contentious one.

One of the most contentious issues, even within the archival science community, is the definition of the term record. For InterPARES 2, a record is "a document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference."[110] The establishment of its characteristics, elements, and attributes is based on archival diplomatics and on the findings of InterPARES 1, which was an archival-process endeavour. The term "record" comes from the Latin "recordari," to remember,[111] but that which is not remembered may be forgotten!

InterPARES 1 stated that five characteristics are required for a digital entity to be a record: stable content and fixed form; embedded action; archival bond; three persons (i.e., author, addressee, writer); and an identifiable administrative and documentary context. For some case studies, particulary Cybercartographic Atlas of Antarctica (CS06)[112] and VanMap (CS24),[113]

110 InterPARES 2, "Glossary and Terminology."

111 Luciana Duranti and Kenneth Thibodeau, "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES," *Archival Science*, vol. 6, no.1 (2006), pp. 13–68.

112 Sherry Xie, *Diplomatic Analysis CS06 Cybercartographic Atlas of Antarctica (revised)* (Vancouver, 2006).

113 Jennifer Douglas, *CS24 Diplomatic Analysis Template Preservation of the City of Vancouver GIS database (VanMap)* (Vancouver, 2006).

which are explicitly designed to allow for data to change and information to be added, this means that they are not or do not contain records in archival terms. To become records, they must be fixed in time and space. IP2, VanMap, and the San Diego Center for Supercomputing have collaboratively designed a research study to determine whether it might be feasible to introduce fixity into the system by changing the system architecture so that each time a layer is updated the layer is saved and set aside. This would allow composite views of VanMap to be assembled for any given date, consisting of layers that had been saved on that date or most recently prior to that date.[114] This is, however, a far from perfect solution, and is both expensive and beyond the capacity of most institutions creating and using these dynamic products. Another approach is to modify and expand the archival definition of a record to reflect the nature of contemporary Internet digital media, such as these two geomatics case studies. The existing debate over records in archival science needs to be broadened and include other disciplines where the term "record" has other definitions and connotations. If this is not done there may never be adequate records of our increasingly participative, interactive digital era. Some of this information may, at best, be preserved but not systematically archived.

One of the most serious problems in this respect is that for most scientists the term "record" means data, databases, and related information, as extensively discussed earlier in this paper. For many archivists, these are not considered records except in very special and limited circumstances, where the concept of "bounded variability"[115] may be applied. This is not simply a matter of semantics. It is a fundamental difference in perspective between creators and preservers, compounded by the emergence in all disciplines of ephemeral interactive information, which exists only in cyberspace. If this problem is not resolved, the increasing volume of interactive, social, and personalized information in the Web 2.0 environment, which does not meet the archival definition of record, may never find its way into archives. This is already happening, and as the interactive, experimental, and dynamic information environment becomes the dominant source of information on many aspects of life in the twenty-first century, we are in danger of losing our cultural heritage.[116] This is particularly the case for many of the IP2 case stud-

114 Evelyn McLellan, *CS24 City of Vancouver Geographic Information System (VanMap)* (Vancouver, 2005). See also the article "From Data to Records: Preserving the Geographic Information System of the City of Vancouver," by Glenn Dingwall, Richard Marciano, Regan Moore, and Evelyn Peters McLellan in this issue of *Archivaria.*

115 McLellan.

116 D.R. Fraser Taylor, Tracey P. Lauriault, and Peter L. Pulsifer, "A Case Study in Geospatial E-science: The Cybercartographic Atlas of Antarctica," *ANAI Seminario Internazionale InterPARES 2 e Seminari Nazionali sul Digitale, Archive e Digitale: Quale Futoro?* (Milan, 2006).

ies (like the Cybercartographic Atlas and VanMap), which are interactive, experiential, and dynamic. Duranti and Thibodeau argue that,

> ... interactions between humans and computer systems, experiences enabled or mediated by experiential systems, and processes which are carried out with at least some degree of spontaneity by dynamic systems are not the residue of action. They are not means of remembering either what was done or what is to be done. In short, they are not records.[117]

This archival position is entirely defensible from the perspective of the theory of diplomatics, but is problematic in many scientific situations, such as for computational data where a model or a simulation is the primary result. The nature of the "record" is changing dramatically and traditional archival science will have to adapt to these changes in both theoretical and practical terms if they are to preserve this new information environment in the archives of the twenty-first century.

Data Archive Initiatives

Even when the problems and challenges of archiving scientific data and digital artifacts are identified, the institutional environment is often not conducive to the systematic action required to address the problem. In Canada, for example, Library and Archives Canada (LAC) is not yet fully ready to systematically archive research data, digital maps, atlases, or the results of complex scientific collaborations, such as genomic projects. Current LAC policies and guidelines for cartographic material primarily address paper maps. The LAC handbook for records and information management, *Managing Cartographic, Architectural and Engineering Records in the Government of Canada* makes only passing reference to digital maps, such as "the National Archives acquires geomatic systems" and "geomatic records include geomatic systems, discs, CD-ROMs and other cartographic material in electronic formats."[118] The Handbook refers the reader to the Canadian Committee on Archival Description's *Rules for Archival Description,* Chapter 5, for information pertaining to standards and practices for cartographic records.[119] These rules

117 Duranti and Thibodeau, p. 59.

118 Library and Archives Canada (LAC), *Managing Cartographic, Architectural and Engineering Records in the Government of Canada* (Ottawa, 2006), http://www.collectionscanada.ca/information-management/002/007002-2050-e.html (accessed 18 January 2007).

119 Canadian Committee on Archival Description, *Rules for Archival Description, Chapter 5 Cartographic Materials* (Ottawa, 2001), http://www.cdncouncilarchives.ca/archdesrules.html (accessed 17 August 2007).

primarily address paper maps while general issues pertaining to digital databases and programs description are covered in Chapter 9: Records in Electronic Form.[120] LAC's *Guidelines for Computer File Types, Interchange Formats and Information Standards*[121] does make reference to some geomatics-specific file types, but adequate guidelines for the kind of multimedia, dynamic, experiential, and multi-sensory digital data emerging in the natural and social sciences still do not exist.

Currently, LAC does not have a digital data archives with the explicit mandate to acquire the results of Internet mapping or scientific endeavours, although a new digital acquisition strategy is under development using a "virtual loading dock." The Social Sciences and Humanities Research Council (SSHRC), the premier Canadian funding agency for social sciences and humanities research projects, explicitly requires that "all research data collected with the use of SSHRC funds must be preserved and made available for use by others within a reasonable period of time."[122] The same policy recommends that researchers ask their university library or data service if it can archive the data, and if it cannot, to ask SSHRC or the Natural Sciences and Engineering Research Council (NSERC) to provide them with a list of possible universities that can assist. The recommended data libraries are not archives but institutional repositories, which are designed to make publications and some data accessible but do not have a mandate to preserve them. In addition, university libraries do not have adequate technical or human resources to archive digital maps, atlases, or complex data. Internationally, there are some social science data archives, such as the UK Data Archive (UKDA), the Council of European Social Science Data Archives, the Inter-University Consortium for Political and Social Research (ICPSR), and some national institutional archives for particular scientific data sets, such as NASA's National Space Science Data Center, or a federated collaborative archive, such as the National Geospatial Digital Archive (NGDA),[123] supported by the Library of Congress, which created the National Digital Information

120 Canadian Committee on Archival Description, *Rules for Archival Description, Chapter 9 Records in Electronic Form* (Ottawa, 2003), http://www.cdncouncilarchives.ca/archdesrules.html. (accessed 17 August 2007).

121 Library and Archives Canada, ed., *Guidelines for Computer File Types, Interchange Formats and Information Standards* (Ottawa, 2006), http://www.collectionscanada.ca/information-management/002/007002-3017-e.html (accessed 18 January 2007).

122 Social Sciences and Humanities Research Council, ed., *Research Data Archiving Policy* (Ottawa, 2002), http://www.sshrc.ca/web/apply/policies/edata_e.asp (accessed 23 August 2007).

123 National Geospatial Digital Archive (NGDA), *Home page* (Washington, DC, 2007), http://www.ngda.org/ (accessed 27 January 2007).

Infrastructure and Preservation Program (NDIIPP)[124]; however, there does not seem to be any natural- or physical-science data nor digital-map archives in any national public archival institution.

There are, however, some initiatives in the long-term preservation of scientific data. GeoConnections is the Government of Canada agency mandated to deliver the Canadian Geospatial Data Infrastructure (CGDI). GeoConnections conducted a study entitled *Archiving, Management and Preservation of Geospatial Data*, which provided a well-rounded analysis of preservation issues in the field of cartography: technological obsolescence; formats; storage technologies; temporal management; and metadata.[125] The study also provides a list of technological preservation solutions with their associated advantages and disadvantages, and a list of proposed institutional and national actions. Phase II of the GeoConnections Program includes archiving as an information management strategy, but these details are still under development. An initiative based out of the Earth Institute at the Columbia University portal for Geospatial Electronic Records includes a number of excellent recommendations regarding the management and preservation of geospatial data, and could potentially assist LAC and GeoConnections with their policies and plans.[126] A new Open Geospatial Consortium Data Preservation Working Group was created in December 2006 to

> ... address technical and institutional challenges posed by data preservation, to interface with other OGC working groups which address technical areas that are affected by the data preservation problem, and to engage in outreach and communication with the preservation and archival information community.[127]

This is a very promising initiative, as the OGC is dedicated to interoperability, open standards, and open specifications, which help overcome many of the issues of platform dependency. The OGC has also done excellent work on the production of the de facto standards of Internet mapping internationally,

124 National Digital Information Infrastructure and Preservation Program, *Digital Preservation: The National Digital Information Infrastructure and Preservation Program* (Washington, DC, 2007), http://www.digitalpreservation.gov/index.html (accessed 27 January 2007).

125 GeoConnections, *Archiving, Management and Preservation of Geospatial Data* (Ottawa, 2005), http://www.geoconnections.org/publications/policyDocs/keyDocs/geospatial_data_mgt_summary_report_20050208_E.pdf (accessed 17 January 2007).

126 Earth Institute at Columbia University, *Geospatial Electronic Records* (New York, 2007), http://www.ciesin.org/ger/index.html (accessed 17 January 2007).

127 Open Geospatial Consortium Data Preservation Working Group, *Preservation Working Group Charter,* (2006), http://www.opengeospatial.org/projects/groups/preservwg (accessed 12 February 2007).

and this working group is dedicated to developing prototypes and test beds with software vendors.

A number of studies, reports, and committees have made high-level recommendations and provided strategies for improving the archiving of digital data in Canada, as they all recognize the poor state of Canada's digital data resources. The SSHRC National Data Archive Consultation report discussed the preservation of data created in the course of publicly-funded research projects.[128] The consultation identified important institutions, infrastructures, management frameworks, and data creators and called for the creation of a national research data archive. The report *Toward a National Digital Information Strategy: Mapping the Current Situation in Canada* indicates that "the stewardship of digital information produced in Canada is disparate and uncoordinated" and in "the area of digital preservation, which involves extremely complex processes at both the organizational and technical levels, comprehensive strategies are not yet being employed. Many feel that much of the digital information being created today will be lost forever."[129] The *Final Report of the National Consultation on Access to Scientific Data,* developed in partnership with the National Research Council Canada (NRC), the Canada Foundation for Innovation (CFI), Canadian Institutes of Health Research, and NSERC expressed concern about "the loss of data, both as national assets and definitive longitudinal baselines for the measurement of changes overtime."[130] This report also provides a comprehensive list of recommendations that include ethics, copyright, human resources and education, reward structures and resources, to name a few, toward the creation of a national digital-data strategy and archive. The CODATA Working Group on Archiving Scientific Data has been holding symposia and workshops on the topic, and the Canadian National Committee for CODATA has been active in documenting and reporting scientific data activities. In December 2006, LAC hosted a National Summit on a Canadian Digital Information Strategy. The challenges of the new Web 2.0 social computing environment, open access, interoperability, and licensing among numerous other topics were discussed at the summit.

128 SSHRC, *Final Report of the SSHRC National Consultation on Research Data Archiving, Building Infrastructure for Access to and Preservation of Research Data* (Ottawa, 2002), www.sshrc.ca/web/about/publications/da_finalreport_e.pdf (accessed 23 August 2007).

129 John MacDonald and Kathleen Shearer, *Toward a National Digital Information Strategy: Mapping the Current Situation in Canada* (Ottawa, 2005), pp. vi, 39, www.collectionscanada.ca/obj/012033/f2/012033-300-e.pdf (accessed 20 January 2007).

130 David F. Strong and Peter B. Leach, *Final Report of the National Consultation on Access to Scientific Data* (Ottawa, 2005), http://ncasrd-cnadrs.scitech.gc.ca/NCASRDReport_e.pdf (accessed 27 August 2007), p. 2.

The report is in its draft form and is expected to be released for public consultation in the autumn of 2007.[131]

The lack of funding to adequately preserve and archive scientific data is a problem. In the United Kingdom "the Economic and Social Research Council will withhold the final 10% of grants if the UK Data Archive cannot confirm that the data generated by the research has been offered to them" but "in general, however, it was felt that agencies were interested more in funding primary research than in providing ongoing support to data archives."[132] In Canada, there are to date no funding schemes in place to support scientists who wish to adequately archive their data, and for many scientists this is not a priority. Archival preservation is often seen as someone else's responsibility.

To date none of the above reports, committees, or recommendations have resulted in the creation of a national science or geomatics data archives, nor have new policies been implemented. Archiving of scientific and geomatics data is technologically complex; however, the greatest obstacles are not technology, techniques, or know-how. The greatest obstacles are the lack of institutional will and the financial resources needed to implement what is already known, and to finance research on unresolved issues. Unfortunately, the situation in Canada is not unique. Many nations and agencies have identified the same problems: few, if any, have implemented the solutions suggested, although the studies discussed in the Canadian context above are steps in the right direction. The following are some high level recommendations distilled from the National Science Foundation's 1995 *Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources*, which encapsulate many of the arguments made in this paper:

1) Data are the lifeblood of science and the key to understanding this and other worlds. As such, data acquired by government or government-funded research endeavours, which meet basic retention criteria, should be considered as critical national resources and must be protected, preserved, and made accessible to all people for all time.
2) The value of scientific data lies in their use. Meaningful access to data is as important as their acquisition and preservation.
3) Explanatory documentation and metadata can ease the use of data.

131 LAC, *Toward a Canadian Digital Information Strategy: National Summit* (Ottawa, 2007), http://www.collectionscanada.ca/cdis/012033-601-e.html (accessed 27 January 2007).

132 Michael Day, "ERPANET and CODATA: The Selection, Appraisal and Retention of Digital Scientific Data: the ERPANET/CODATA workshop," *Ariadne* 39 (2003), http://www.ariadne.ac.uk/issue39/erpanet-rpt/ (accessed 17 August 2007).

4) A data archives should be extensible, durable, readily accessible, and affordable.
5) Distributed archives are recommended and data should be managed wherever possible by their creators.
6) Long-term management of data needs to be planned into the process at the point of creation.

Conclusion

The research carried out by IP2 on these scientific and e-government case studies, and the data portals of General Study 10 reveals a number of key issues in the preservation of scientific data. Data are the bread and butter of science; they help form baselines upon which we base decisions and plan. The longer the timeline of the data set, the more robust the record of an event, experiment, or simulation. Data on their own, however, are inadequate. Scientists need metadata to make fit-for-use decisions and, within metadata, they need respect for specific data quality parameters that relate to accuracy, reliability, and authenticity; they will not trust data without adequate documentation. Errors are implicit in any data set, simulation, model, or experiment, therefore, the margin of error needs to be explicit to inform scientists on inherent limitations.

Science is a heterogeneous discipline, and each field and subfield has its own culture, methods, quality measures, and ways to explain what they do. Formal ontologies are a method to help mediate the myriad metadata standards and facilitate the production of meaningful ways to represent the world and preserve the data. Data sets are often accessed via data portals that are research, community, or reference collections, and are organized into distributed, collected, or unified cataloguing systems. Furthermore, data portals reflect the policies, funding agencies, and the technologies chosen by the organizations that manage them. Organizational, technological, metadata, and data quality aspects affect appraisal decisions and provide challenges for archivists.

Science is a collaborative endeavour that is premised on the notion of knowledge sharing, dissemination, reproducibility, verification, and the possibility that new methods will yield new results from old data. Therefore, there is an argument to be made that publicly-funded collections of data should be made available to the citizens who paid for them and to future generations for the advancement of knowledge. For a scientist, data and related scientific information are records. Archivists dealing with scientific data and records must come to terms with this challenge, or their relevance and utility to the sciences will be adversely affected. This issue is not new but takes on new importance in an increasingly digital data world, particularly in an environment of experiential, dynamic, and interactive scientific data.

The IP2 research showed that interoperability is a problem with the rapidly increasing number of digital databases that need to interact, one that challenges knowledge integration. The Cybercartographic Atlas of Antarctica was faced with the challenge of using information from different databases in different countries and, in order to do this, adopted an open source and open standards approach. This decision was taken primarily for production reasons but has had beneficial effects in archiving and preservation terms as it helps overcome the problem of technological obsolescence. The IP2 case studies demonstrate that a lack of interoperability can lead to having data that cannot be archived in the same form the creator intended. Indeed, it can be argued that interoperability is a key element in archiving all digital data and that an open source standards and specifications approach should be a facet of any archival strategy.

> The way ahead lies in an innovative combination of both approaches. Geospatial E-science must give much more attention to creating records which can be preserved and the preservers must listen to other perspectives on the definition of the term "record."[133]

This applies more widely to the sciences as a whole.

For scientific disciplines, trust will continue to rest on specific norms of scientific work. Additionally, standards for managing data and metadata in digital media that are controlled by software systems and accessed through communication technologies and proprietary hardware, including data validation, processing, compilation, and aggregation, will need to be developed, tested, and put into place. Trusted repositories, whose data is kept reliable, accurate, and authentic over time, will need to be established, managed, and funded on a continuing basis. The problems are on three levels: organizational stability, data and metadata management processes, and technological handshaking across generations.

Established archival repositories, data-loading docks with access mechanisms that are mandated (and funded) to guarantee the continuing availability of scientific data records and information that support administrative, legal, and historical research are needed. Although there are digital repositories for social science data, true digital scientific data archives are few and far between. The IP2 General Study on data portals demonstrated that there are numerous excellent initiatives in place to make data discoverable and accessible. However, few of these data portals archive their data. The few portals that are government funded in the US and simultaneously housed in government departments do have preservation as a mandate or are considered to be

133 Taylor, Lauriault, and Pulsifer.

government archives; most portals, however, do not have this type of financial or institutional stability. At risk in particular are the repositories that are distributed and leave issues of data quality to the data custodians or creators. Therefore, much government funded science is not enveloped in any data preservation or archiving processes. This is quite troubling, considering the investment taxpayers have made in these endeavours, let alone the loss in knowledge disseminating, and building opportunities.

In Canada, there is much discussion about the archiving of digital data, and some organizations have excellent guidelines in place; yet there is no one agency or archive that currently acquires publicly-funded data. Some specialist repositories exist to support the continuing information needs of particular communities of interest that often span the organizational boundaries of established administrative repositories. Joint or collaborative projects of research, interdisciplinary knowledge sharing, market information for competitive advantage, or grass-roots environmental monitoring initiatives require new structures and archival arrangements to manage their accumulating knowledge, information, and data for current uses and future reuse. Some initiatives show promise, such as the Open Geospatial Consortium Data Preservation Working Group and the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP), as do the portals that are designated science data archives or have preservation mandates, and GRID Supercomputing projects.

In deciding what data is archived, it is perhaps a cliché to suggest that in science (and perhaps in other disciplines) archiving is too important to be left only to archivists. All stakeholders, including the scientists who create the information, research managers, major user groups, and of course the archivists, should be involved in the appraisal decisions on what is to be archived, and by whom. This appraisal should be an ongoing process from the point of creation and is best carried out in a project-specific fashion, in collaboration with those most knowledgeable about the data. Some key appraisal questions remain. Are the data unique, accessible, and accompanied by adequate metadata? Are the observations reproducible? What is the quality of the data? Has the science involved been subject to peer review? Can the data collections in a portal be considered appraised by the scientific community already? Can the archivist risk preserving only subsets of large data sets, or will this cause problems with future statistical analyses?

In today's increasingly ephemeral and interactive digital world of social computing, much information may be archived by individual creators rather than through formal archiving institutions. Many members of the current Web 2.0 generation carry their "archives" on their mobile devices or on memory keys worn as jewelry around their necks. Many scientists do the same with their scientific data. Will the twenty-first-century archives capture and "record" these data or will they continue to be lost, as is unfortunately very

much the current situation? Can archivists build on the energy in place in the data creation and maintenance and data portals and extend these activities with archival policies, techniques, and technologies? Can archivists work collaboratively with initiatives like the OGC and extend some of that capacity into the archives? Will geospatial data infrastructure initiatives include preservation as a new component? What is the best method to appraise data quality? And, are formal ontologies a possible solution? Will today's data be a part of tomorrow's research? The answers depend on how all concerned respond to the challenges and possibilities discussed in this paper.

Appendix 1: Selected InterPARES 2 Case Studies

ID	Selected IP2 Case Study	IP2 Focus	Description	Observational Data	Computational Data
CS06	Cybercartographic Atlas of Antarctica (Lauriault and Hackett 2005)	Science	On-line interactive and dynamic, open standards, interoperable multi-media, multi-sensory, multi-modal Atlas that renders distributed data from myriad scientific organizations.	The atlas renders distributed observational data from myriad sources in real time (e.g., film, satellite images, tabular data, sounds, etc.)	Data are rendered/ refined into maps, charts, tables by the Nunaliit Atlas Framework.
CS08	Mars Global Surveyor Data Records in the Planetary Data System: A Case Study (Underwood 2005)	Science	NASA Mars Global Surveyor mission data records at the Planetary Data System (PDS) Space Science Data Archive.	Level 0 and refined planetary spacecraft mission data stored in a database with attributes and metadata.	Software used to access or visualize the data.
CS14	Coalescent Communities in Arizona (O'Meara, Pearce-Moses, and Preston 2004)	Science	Archaeological Records of the American Southwest rendered in a Geographical Information System (GIS).	Tabulated raw data collected from archeological digs.	Raw data are rendered/refined into maps.

ID	Selected IP2 Case Study	IP2 Focus	Description	Observational Data	Computational Data
CS18	Computerization of Alsace-Moselle's Land Registry (Blanchette 2004)	Government	Electronic registry including digital transcription of 40 000 existing paper registries and new database entries individually signed by a judge using a PKI infrastructure combining biometric access and digital signatures.	Database of digitized paper (land registries and attributes).	Data and their attributes are accessed via a database using PKI technology.
CS19	Authenticating Engineering Objects for Digital Preservation (Hawkins 2005)	Science	Examines through an engineering experiment the authentication of digital model (CAD) records using a content/message/semantic-based methodology rather than media, bitcount, or static provenancial attribute-based authentication.	CAD solid model files used in the design and manufacturing of mechanical piece-part assemblies.	The files are rendered and stored in a proprietary system.
CS24	City of Vancouver Geographic Information System (VanMap) (McLellan 2005)	Government	An enterprise web-based map system maintained by the City of Vancouver's Information Technology Department.	Land use, social statistics, city infrastructure data both internally collected and acquired from external sources, etc.	Enterprise GIS renders data into maps, charts, tables, etc.
CS26	Most Satellite Mission: Preservation of Space Telescope Data (Ballaux 2005)	Science	Repository of the Microvariability & Oscillations of Stars (MOST) satellite mission data of Canada's first space telescope.	Raw satellite data and their refined counterparts.	

Appendix 2: Selected InterPARES 2 General Study 10 (GS10) Scientific Data Portals

IP2 ID	Portals	Description	Data Sets
IP2SF1	British Atmospheric Data Centre (BADC)	Assists UK researchers to locate access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by Natural Environment Research Council (NERC) projects.	Data sets produced by NERC-funded projects; some third-party data sets that are required by a large section of the UK atmospheric research community and are most efficiently made available through one location.
IP2SF2	NASA Life Sciences Archive	Information and data from space flight experiments funded by the National Aeronautics and Space Administration (NASA).	Investigations from 1961 (Mercury Project) through current missions (International Space Station and Shuttle) involving human, plant, and animal studies. Collection of data describing, in considerable detail, biology experiments. The nature of the data is highly varied and spans many life-science disciplines.
IP2SF3	University of Washington: Electrical Engineering Circuits Archive (EECA)	Collection of a wide range of circuit designs for engineering students and professionals.	Circuit designs and related information, circuits, data sheets, models, microprocessors, text files, software, seading capacitors are resistor codes.
IP2SF4	Cambridge Crystallographic Data Centre	Curator of the world's output of small molecule crystal structures, and to undertake activities in support of related academic and industrial research.	Mean molecular dimensions, studies of substituent effects, statistical and numerical data analysis techniques, structure correlation and reaction pathways, conformational analysis, Hydrogen bond geometry and directionality, weak hydrogen bonds, non-bonded interactions not mediated by hydrogen, crystal engineering, crystallographic symmetry, and molecular symmetry.

IP2 ID	Portals	Description	Data Sets
IP2SF5	IU (Indiana University) Bio Archive	Public access to freely available information, primarily in the field of molecular biology.	Molecular data, public software, biology news and documents, Drosophila research data.
IP2SF6	Computational Chemistry Archives/ Computational Chemistry List	Fosters communication within the worldwide community of researchers involved in chemistry-focused computation. It is a list exploder.	The list includes reporting bugs in chemistry software and possible work, new chemistry software announcements, announcements of computational chemistry-related workshops and symposia, new methods and techniques in computational chemistry, solving computational chemistry problems, programs, and utilities. List messages are exploded and their content organized.
IP2SF7	The FMRI Data Center (fMRIDC) [Functional MRI]	Providing access to a common FMRI data set to develop and evaluate methods, confirm hypotheses and perform meta-analyses, and increasing the number of cognitive neuroscientists who can examine, consider, analyze, and assess the brain imaging data that have already been collected and published.	Peer-reviewed, published FMRI (functional magnetic resonance imaging) studies, the pre-processed images, and necessary technical descriptions. Raw, reconstructed image volumes from the scanner, pre-processed images used for statistical analyses and detailed descriptions of the image processing steps that were applied, high-resolution anatomical images from all subjects (e.g., SPGR, MPRAGE, etc.), image volumes of final statistical results for each subject and statistical group maps.

IP2 ID	Portals	Description	Data Sets
IP2SF8	NIST (National Institute of Standards and Technology) StRD Statistical Reference Data Sets (Data set Archives)	Improves the accuracy of statistical software by providing reference data sets with certified computational results that enable the objective evaluation of statistical software.	Generated data sets designed to challenge specific computations, Wampler data sets for testing linear-regression algorithms and the Simon & Lesage data sets for testing analysis of variance algorithms. Real-world data include challenging data sets such as the Longley data for linear regression, and more benign data sets such as the Daniel & Wood data for nonlinear regression.
IP2SF9	Animal Cognition Laboratory, Department of Physics, University of Georgia Data Archive	Data sets contributed and maintained by faculty and students in the Department of Psychology at the University of Georgia.	Raw data from experiments conducted in the Animal Cognition Lab.
IP2SF10	World Data Center for Solar Terrestrial Physics	Access to solar, geophysical, and related environmental data.	It includes International Geophysical Year (IGY) data, solar and interplanetary phenomena, ionospheric phenomena, flare-associated events, geomagnetic variations, and cosmic rays.
IP2SF13	OBIS-SEAMAP (Ocean Biogeographic Information System – Spatial Ecological Analysis of Megavertebrate Populations)	To augment understanding of the distribution and the ecology of marine mammals, birds, and turtles.	Data sets include marine mammal, seabird, and sea turtle distribution and abundance.

IP2 ID	Portals	Description	Data Sets
IP2SF14	Canadian Institute for Health Information (CIHI)	Coordinate the development and maintenance of a comprehensive and integrated approach to health information, and coordinate the provision of accurate and timely data and information.	The data consist of health data provider demographics, health human resources, health spending, and health services and health issues, drug, and hospital-generated data.
IP2SF15	Canadian Geospatial Data Infrastructure (CGDI) Access Portal	To deliver Canada's Geospatial Data to Canadians via the Internet.	Geospatial data such as DEM, orthophotos, air photos, satellite and radar imagery, data services, discovery metadata, documents, Internet maps, and atlases.
IP2SF16	The National Cancer Registry (NCR) Ireland	Identify, collect, classify, record, store, and analyze information relating to the incidence and prevalence of cancer to facilitate use of the data in approved research, planning, and management of services.	Cancer-related tumors, cancer incidences in relation to each newly diagnosed individual cancer patient and in relation to each tumor which occurs.
IP2SF17	National Institutes of Health (NIH)	Supports the research of non-Federal scientists in universities, medical schools, hospitals, and research institutions; helping in the training of research investigators; and fostering communication of medical and health sciences information.	Scientific data in peer-reviewed papers, reports and the results of state funded research projects and NIH laboratories.

IP2 ID	Portals	Description	Data Sets
IP2SF18	Statistics Canada	The National Statistical Agency	Under the *Statistics Act*, Statistics Canada is required to collect, compile, analyze, abstract, and publish statistical information relating to the commercial, industrial, financial, social, economic, and general activities and conditions of the people of Canada.
IP2SF19	National Virtual Observatory (NVO)	Makes astronomical science data discovery in the world easy to access, using a simple set of Web interfaces.	The NVO does not collect any data of its own; instead, it provides the resources to let users search and analyze data that already exists.
IP2SF20	TeraGrid Data Collections	Is an open scientific-discovery infrastructure combining leadership class resources at eight partner sites to create an integrated, persistent computational resource.	Data relating to natural sciences; over 100 discipline-specific databases and multimedia data for heritage sites.
IP2SF21	Joint Center for Structural Genomics (JCSG)	Establishes a robust and scalable protein structure determination pipeline to form the foundation for a large-scale effective production center for structural genomics.	Genomics research, and related infrastructure and computing resources: Bioinformatics; Crystallomics and Structure Determination data.
IP2SF22	San Diego Supercomputing Centre (SDSC)	Provides high-performance hardware technologies, integrative software technologies, and deep inter-disciplinary expertise.	Sensor data, visual simulations and libraries of highly refined and analyzed data.

IP2 ID	Portals	Description	Data Sets
IP2SF23	Long Term Ecological Research (LTER)	Provides the scientific community, policy makers, and society with the knowledge and predictive understanding necessary to conserve, protect, and manage the nation's ecosystems, biodiversity, and services.	Data of Holocene barrier island geology; salt marsh ecology, geology, and hydrology; ecology/evolution of insular vertebrates; primary/secondary succession; life-form modelling of succession; on-line maps, photos, webcams; physical; biological, images and geographic data, software, LTER Network Data; and models.
IP2SF24	Southern California Earthquake Center (SCEC)	Maintains an easily-accessible, well-organized, high-quality, searchable archive of earthquake data for research in seismology and earthquake engineering.	Data sets include seismic waveforms (mostly passive source seismic data collected by broad-band, strong-motion and analog instruments). LARSE I and II seismic survey, data sets from the portable deployments following the Landers and Northridge earthquakes, and data from the Anza network and SCEC borehole stations. Non-seismic data includes "survey-mode" precise GPS measurements made in southern California by various universities. The GPS data archive consists of raw GPS data, RINEX files, indices to the RINEX files, log sheets, and site descriptions.
IP2SF25	International Comprehensive Ocean Atmosphere Data Set (ICOADS)	Global surface marine data from the late 18th century to date, which have been assembled, quality controlled for the international research community. Data produced by three national organizations.	Surface marine reports from ships, buoys, and other platform types. Each report contains individual observations of meteorological and oceanographic variables, such as sea surface and air temperatures, wind, pressure, humidity, and cloudiness; monthly summary statistics.

IP2 ID	Portals	Description	Data Sets
IP2SF26	National Geophysical Data Center (NGDC - NOAA)	Access to global environmental data from satellites and other sources to promote, protect, and enhance the nation's economy, security, environment, and quality of life.	Geophysical data describing the solid earth, marine, and solar-terrestrial environment, as well as earth observations from space.
IP2SF27	Antarctic Digital Database (ADD)	Seamless digital maps of the Antarctic.	1:200,000/1:250,000 maps and collaborative topographic database compiled from a variety of Antarctic map and satellite image sources.
IP2SF28	National Snow and Ice Data Center (NSIDC), NASA	Manages cryospheric science data and disseminating information in order to advance Earth system research.	Snow, ice, and glaciological data from satellite images, remote-sensing instruments, ground measurements, and data models.
IP2SF29	US Antarctic Resource Center (USARC)	Reliable scientific information to describe and understand the Earth; minimize loss of life and property from natural disasters; manage water, biological, energy, and mineral resources; and enhance and protect quality of life.	Maps, orthophotos, satellite imagery, point data, Digital Elevation Models, Digital Raster Graphics, and areal photography,
IP2SF30	British Antarctic Survey – Antarctic Environmental Data Centre	Program of scientific research to sustain for the UK an active and influential regional presence and leadership role in Antarctic affairs.	Antarctic research results, climate modelling and predictions. Oceanic, environmental, atmospheric, geoscience, water, and earth observation data from a variety of sensors and in many forms.

IP2 ID	Portals	Description	Data Sets
IP2SF32	Global Change Master Directory – Global Change Data Center	Enables the scientific community to discover and access earth-science data and services through distributed, integrated information technology systems.	Data inform climate change research and data sets include models, instruments, and services. It is a work space as well as a place to store these working operational models. Disciplines include atmospheric science, oceanography, ecology, geology, hydrology, and human dimensions of climate change.
IP2SF35	Community Data Portal at NCAR	State-of-the-art data portal with a broad spectrum of functionality.	CDP catalogs observational and computer simulation data sets, sources of data are related to sciences in the areas of: oceanic, atmospheric, space weather, and turbulence.
IP2SF36	Earth Systems Grid (ESG) portal	Seamless and powerful environment that enables the next generation of climate research through a combination of Grid technologies and emerging community technology, distributed federations of supercomputers, and large-scale data and analysis servers.	Climate-change models include high-resolution, long-duration simulations, data simulations, derived from UCAR/NCAR projects (particularly the Community Climate System Model (CCSM), Parallel Climate Model (PCM), and the Intergovernmental Panel on Climate Change (IPCC). Also physical data sets that are generated directly by climate simulations.
IP2SF37	USGS Data Portals – GEO-DATA Explorer (GEODE)	Access, view, and download information from geo-spatial databases containing a broad spectrum of data produced by the USGS and other government agencies.	Geologic discipline's programs including coastal and marine geology, earth surface dynamics, earthquake hazards, integrated natural resource sciences, mineral resources, National Cooperative Geologic Mapping, and volcano hazards. Also scientific and energy related data.