# "Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences"
# *Archivaria 64* (Fall 2007)

## Supplementary Data Tables

This document comprises tables containing data gathered during the case studies conducted by the InterPARES 2 project (http://interpares.org/ip2) to accompany the article "Today's Data are Part of Tomorrow's Research: Archival Issues in the Sciences," by Tracey P. Lauriault, Barbara L. Craig, D. R. Fraser Taylor, and Peter L. Pulsifer, which appears in *Archivaria* 64 (Fall 2007), pp. 123-79. Details of the case studies themselves can be found on the InterPARES 2 website at http://www.interpares.org/ip2/ip2_case_studies.cfm (accessed 6 September 2007).

### Table 1. InterPARES 2 Selected Case Study Authenticity Responses.

| ID | IP2 Case Study Name | Authenticity in InterPARES 2 Selected Case Study Reports |
|---|---|---|
| CS06 | Cybercartographic Atlas of Antarctica (Lauriault and Hackett 2005) | The creator uses data that have adequate metadata and that are from reliable trustworthy source and each module is accompanied by the creator's metadata, and access to the content is secured. "The online CAA production environment is protected by security measures such as physical security, and password protection. Access to the CAA itself is restricted to the CAA's technical specialists". "The CAA will have its own domain name, and a trademark with branding. GCRC researchers oversee the modules, providing additional checks and balances". |
| CS08 | Mars Global Surveyor Data Records in the Planetary Data System: A Case Study (Underwood 2005) | "Due to the emphasis on completeness and reliability of the planetary science data records, the peer review, role-based authentication of access to archived data products, and data integrity checks, that the scientific data records are maintained authentic". |
| CS14 | Coalescent Communities in Arizona (O'Meara, Pearce-Moses and Preston 2004) | "The creator discerns that the records are authentic, to an acceptable degree for the research that is being conducted. There is more concern over the reliability and accuracy of the records than the authenticity. Authenticity is assumed to a reasonable degree since records are coming from a state repository or a data set from a researcher that is trusted as a professional that maintains their information." |

| | | |
|---|---|---|
| CS18 | Computerization of Alsace-Moselle's Land Registry (Blanchette 2004) | "Elaborate technical and procedural measures directed at ensuring the integrity and authorship of the land registry data, it has no reasons to believe that the authenticity of the information it provides will prove more questionable than that provided by the paper-based registry. The computerization process may even improve in certain aspects the reliability of the land registry, by eliminating the need for transcription of complex real estate information". "As a last note, outside users access the computerized land registry using SSL-secured communication, thus guaranteeing that the information cannot be modified in transit between their workstation and the server" |
| CS19 | Authenticating Engineering Objects for Digital Preservation (Hawkins 2005) | Creators do not believe their records are authentic and "they realize that there is no assurance system" and faith is placed in the CAD software only in its original form. |
| CS24 | City of Vancouver Geographic Information System (VanMap) (McLellan 2005) | VanMap "will guarantee that the data are as reliable as the data provided by the original source. Only a limited number of people within the departments are able to update data, staff entering the data are well trained, and the data entry formats are strictly controlled: according to one interview subject". "In most cases, the departments consider themselves the authoritative sources of the information. For example, decisions about zoning are made by City Council through amendments to the Zoning By-law, but the responsibility for creating the zoning maps based on these amendments lies with CSG, which assumes responsibility for the accuracy of the work. CSG considers itself the authoritative source for information about licenses, and takes steps to ensure that very few staff is able to change the data. Preventing unauthorized access to License+ is one of the reasons for sending an extract to a separate server for inclusion in VanMap". |
| CS26 | Most Satellite Mission: Preservation of Space Telescope Data (Ballaux 2005) | "There is no reason for the creator to assume that the digital entities are not authentic". The process to acquire, manipulate and store these data are mostly machine done." |

# Table 2: Accuracy Statements from the InterPARES 2 Case Studies in the Sciences and Geomatics

| ID | Case Study Name | Accuracy |
|---|---|---|
| CS06 | Cybercartographic Atlas of Antarctica (Lauriault and Hackett 2005) | "Data are acquired from authoritative sources and are peer-reviewed". "Quality measures are dependant on the type of data and their function (e.g. the acceptable margin of error for the precise location and size of a particular ice flow to inform tourist ships is smaller than fish counts to inform fisheries and ecological modeling). In addition, each scientific domain is governed by their particular data quality standards, measures and assurances and these are included in the metadata." "Disciplines and professions represented in this project also have established procedures and practices to ensure data quality, reliability and authenticity." "Some metadata will be embedded into information objects; additional metadata will be linked to the objects at the module level." |
| CS08 | Mars Global Surveyor Data Records in the Planetary Data System: A Case Study (Underwood 2005) | "A Project Data Management Plan (PDMP) is required by NASA for all new projects. This plan provides a general description of the project data processing, cataloging, and communication plan [JPL]." "The *PDS Data Preparation Workbook* (DPW) serves as a guide for the organization and preparation of data sets intended for submission to the Planetary Data System (PDS) [JPL1995a]." "The Archive Policy and Data Transfer Plan (APDTP) provides a detailed description of the production and delivery plans for archive products for a project. A Data Product Software Interface Specification (SIS) is a document that describes the format and size of the individual data products. All data incorporated into the PDS archives must undergo a peer review. The purpose of the review is to determine that:<br>• The data is accurate, complete and reliable.<br>• The data are suitable for archiving.<br>• The PDS standards have been followed. [JPL 2003]"<br>"The primary validation tool of the PDS is the Volume Verifier. The Central Node data engineers run this program on each product delivered from a project. It validates the format and content of all product labels, and validates the integrity of data files using checksums." |

| | | |
|---|---|---|
| CS14 | Coalescent Communities in Arizona (O'Meara, Pearce-Moses and Preston 2004) | "Throughout the process of creating the database, spot checks are being conducted on the data set to ensure that data is being entered accurately. In general, it is difficult to deduce if the record is authentic in some sense of the word. Archaeological data that is used is usually coming from trusted sources where the records are held in a repository. Other times, the records are coming directly from the researcher's personal computer. There is usually a preliminary audit stage where the GIS Specialist checks the data as a whole to see if it is generally reliable and authentic as a whole data set. In addition, the volunteer checks these aspects of the data set in a more thorough manner with his expertise in the field being a strong factor in effectively probing the data set. Archaeologists know that their data is not 100% reliable to fact, due to the nature of the archaeological record, so there is a degree of reliability that needs to be met before the data is considered "usable". |
| CS18 | Computerization of Alsace-Moselle's Land Registry (Blanchette 2004) | "Legal: the inscriptions within the land registry have a specific evidential value (presumption of correctness), while the ordinances they are based on have the status of "authentic" acts. Because the computerized land registry continues to rely on the same legal professionals — the judges du livre foncier — and because French evidence law as already been reformed to account for the evidential value of electronic information, the computerized land registry will continue to hold information admissible in court as evidence."<br>"Professional: the reliability of the land registry results from the intimate familiarity of the clerks and judges with the paper system. The GILFAM has involved those professionals into the design process of the computerized land registry, and has put tremendous effort into what it calls the "management of change", in order to insure that future operators and users of the system will be maximally comfortable with its new configuration." Technically for inscriptions, the system includes the verification of scanned images, PKI technology for signatures of judges on ordinances, transcription checks, "It expects to control these rates to below 0.1% for the digitization process, and below 0.5% for the transcription." New digital registries "within the database itself, data is protected from malicious modifications only through the access control mechanisms in place, and protected from accidental modification only through the general soundness of the software architecture." The system includes designed verification systems for inscription and ordinances, which "Using a comparison mechanism between the inscriptions and the signed ordonnances provides the necessary flexibility for system evolution, while retaining the high integrity standard of digital signature". |
| CS19 | Authenticating Engineering Objects for Digital Preservation (Hawkins 2005) | "The designer follows quality rules regarding the way that they construct the solid model. There is a quality guide for this. Also the creator has the option to assess the geometric quality of the model with a checker. Typically CAD systems have rigorous model checks to make sure that the model construction step performed by the creator does not create bad geometry. The design/manufacturing engineer does nothing to ensure the reliability and authenticity of the digital entities other than the quality checks. Support persons ensure that the creator is using the right version of the software". |
| CS24 | City of Vancouver Geographic Information System (VanMap) (McLellan 2005) | "Data quality is not the responsibility of the VanMap developers but rather of the originating departments. The Engineering Services representative on the VanMap Team stated that "each data layer is signed off by the data owner. I don't proceed with moving anything into VanMap unless I have a commitment from the data owner that the data is as complete as we can |

| | | |
|---|---|---|
| | | make it, and as accurate as we can make it." 'These commitments are not formalized, however, but are either agreed to verbally or through e-mail (the City has no e-mail management policy). The CSG representative noted that there's an element of trust between the VanMap Team and the departments – the Team trusts that when a department says the data are reliable, they are in fact reliable. As she put it, "we all work for the City." Data are trusted but before major works at a site are conducted there some ground truthing exploration are done. "On the other hand, the business unit responsible for supplying certain data may feel confident that the data are accurate. For example, the Traffic Management Division tends to view the traffic count data it supplies to VanMap as being accurate and reliable. Staff may be more reluctant to rely on data supplied by other departments" "The VanMap Team will guarantee that the information that appears in VanMap is as good as the original source. In other words, when they are added to VanMap the data are not altered in such a manner as to affect their accuracy and reliability" |
| CS26 | Most Satellite Mission: Preservation of Space Telescope Data (Ballaux 2005) | "Throughout the process of transmission of data, some check sums are in place to ensure that the data have not been changed. This is a technical issue, and has to be done for all raw data files to ensure that they are reliable. In addition, the instrument scientist controls the data in an intellectual way, by looking to individual files or sequences of files. From the moment these checks have been done, the MOST researchers consider the data as good and reliable." However, the researchers acknowledge that files with corrupt raw data from the satellite are also used. In reducing the data, these corrupt or false data are eliminated. The fact that these data are false or corrupt has no impact on the outcome of the reductions (one of the end results of the scientific research). Whether or not these false or corrupt data are included in the calculations for the reduction, does not affect the reliability of the outcome. Accuracy and reliability is an important issue for the FITS files as well. In the process of creating these FITS files, there are various checks to assure that the information that is put in, is good. If errors occur, the researchers typically will examine the problem and recreate the FITS files. For instance, a magnetic field value is added in the FITS files. This value is based on the position of the satellite at a particular time, and is calculated on the basis of a model. The model that was used only went up to 1 January 2005. In the beginning of January, all magnetic field values were 0. After the problem was identified, the FITS files were recreated on the basis of a new magnetic field model. A researcher notes that it is possible that in the future, a value has to be added or corrected in the FITS file. In this sense, the researchers don't use the words accurate and reliable in an absolute sense." |

Table 3: Metadata in Selected Science and Geomatics InterPARES 2 Case Studies

| ID | IP2 Case Study Names | Metadata in Selected IP2 Case Studies |
|---|---|---|
| CS06 | Cybercartographic Atlas of Antarctica (Lauriault and Hackett 2005) | "At minimum, the atlas and each module will have an ISO19115 record. In part, this ISO record will be built from a standard 'resources' or 'references' section in each module. This resources section contains information related to a module (e.g. literature citations, sources for multimedia etc.)" (Pulsifer 2006). The Atlas contains many multimedia objects and a GCRC Multimedia Metadata standard was developed by Y. Zhou, *Profiling and Visualizing Metadata for Multimedia Information in a Geospatial Portal*. Some of the data accessed by the atlas will include Federal Geographic Data Committee (FGDC) metadata standard and British Antarctic Survey DIF or Directory Interchange Format for the Antarctic Digital Database. |
| CS08 | Mars Global Surveyor Data Records in the Planetary Data System: A Case Study (Underwood 2005) | "The *Planetary Science Data Dictionary* (PSDD) being used in the creation, maintenance, use and preservation of the Planetary Data System. The PSDD contains definitions of the standard data element names and objects." "The Planetary Science Data Dictionary is a NASA, institutional standard for Planetary Science Metadata. The PDS procedures for assigning standardized names to data elements follow closely the NBS Guide on Data Entity Naming Conventions, [NBS]." |
| CS14 | Coalescent Communities in Arizona (O'Meara, Pearce-Moses and Preston 2004) | There are no metadata. There is interested in using ArcCatalogue. The only metadata at the moment indicates from what source (publication, repository, website, database, etc.) the data were retrieved. |
| CS18 | Computerization of Alsace-Moselle's Land Registry (Blanchette 2004) | "Both ordonnances and inscriptions are captured through custom applications. The scanned images of the register were captured once at the onset of the computerization process". There are no metadata. |
| CS19 | Authenticating Engineering Objects for Digital Preservation (Hawkins 2005) | "Currently ISO 10303 is most important standard. All of our drawings conform semantically to the ANSI Y-14.5 tolerance standard. We also have corporate standards for solid-model and drawing metadata. We also abide by the corporate guideline to store CAD native model, STEP model, and .TIFF image of generated drawing together as an Aggregate that cannot be separated. We do not know to what degree we can assure the immutability (fixity?) of this aggregate over time as we are quite sure that these three digital objects are related to each other through relationships in the product data management system". |
| CS24 | City of Vancouver Geographic Information System (VanMap) (McLellan 2005) | The organization and schema are dictated by the nature of the activities used to produce the data. For example, processing a building permit application requires attribute data about property address along with other details pertaining to the permit application and approval process be recorded and maintained together. "When the data are extracted for inclusion in VanMap they are kept in this logical grouping, and the reports generated in VanMap are designed to provide access to the data in the same logical grouping. The geospatial information is organized by or linked to location information, since the purpose of VanMap is to provide information linked to the geography and physical features of the City." "The metadata are based on what the VanMap Team thinks will be useful information for the end user". |
| CS26 | Most Satellite Mission: | The metadata schema used was created by the MOST researchers, and is |

| | | |
|---|---|---|
| | Preservation of Space Telescope Data (Ballaux 2005) | specific for the data/files that are created in the MOST project. "The metadata refer to information such as orbital parameters, observational parameters, telemetry information, and target image information. Some of these metadata/descriptive fields in the FITS files are mandatory, due to the file format. In general, no metadata standards are used, the MOST researchers created their own scheme of important descriptive fields." |

.

# Table 4: InterPARES2 Case Study Respondent Record statements and Diplomatic Analysis Conclusions

| ID | IP2 Case Study Names | What Case Study Respondents consider to be Records | Conclusions from the Case Study Diplomatic Analysis |
|---|---|---|---|
| CS06 | Cybercartographic Atlas of Antarctica (Lauriault and Hackett 2005) | "Of particular importance to the long-term viability of the CAA are the XML-tagged content modules created by the content creators. These are considered the "master" content element. They are processed via a compiler to make them web-ready. Should the technology platform of the CAA change, the content of the Atlas would be re-built by re-accessing the XML content modules and processing them anew through a new compiler. While this method will not protect all information objects included in the CAA. (I.e. sound, video, Flash, etc.), it should facilitate forward migration of the essential content, presentation information and intended functionality. Proprietary problems remain with some multimedia formats used in the CAA." | "Atlas only satisfy some of the conditions (e.g., the persons concurring in the process of its creation and some contexts) that are necessary to be considered as a record. The fact that its content (with fixed documentary form unable to be decided) is subject to continuous changing/updating, it does not participate in one action, and it does not possess an archival bond makes the Atlas only partially satisfies the definition of record. This diplomatic analysis thus concludes that the Atlas is not a record. This analysis also reveals the characteristics of the CAA as a publication.26 As stated in the above discussion regarding archive bond, the Atlas, upon completion, becomes a self-contained entity, standing on its own and does not require any other information to make it to be understood. It is presented to the public and publicly accessible, and it will have its own domain name, and a trademark with branding. Like any other publications, the messages it contains are self-explanatory and the meaning it conveys is complete. Even in a dynamic and interactive environment, every instantiation of the assembled data or every display responding to user |

| | | | |
|---|---|---|---|
| | | | inquiries is autonomous. Its publication status remains as long as it is hosted online and consulted as an Atlas" (Xie 2006). |
| CS08 | Mars Global Surveyor Data Records in the Planetary Data System: A Case Study (Underwood 2005) | The measurements received from the instruments at the mission ground control system are considered to be "experimental data records (EDRs)" and these include "data products consist of EDRs, SPICE kernels, Standard Data Products, and Special Data Products (SPDPs). SPDPs are defined as those science data products produced during the course of science analysis." | Not Available |
| CS14 | Coalescent Communities in Arizona (O'Meara, Pearce-Moses and Preston 2004) | "The creator believes that the GIS is a record once the findings and data in it are published through a journal article or monograph. In order to come to the point of publication, the creator has to "come to a meaningful anthropological conclusion about something…when I write that up and send it away and hopefully it gets published that would be my definitive answer". And "the creator treats the various versions of the CC database as records when they are purposely set aside during the course of business. They fulfill a need so that if the new data that has been added to the database is inaccurate or corrupt, there is a previous version that can be retrieved and used." | "According to the above analysis the authoritative record is the GIS itself" (Catto and Preston 2006). |
| CS18 | Computerization of Alsace-Moselle's Land Registry (Blanchette 2004) | "For the GILFAM, the ordonnances, the inscriptions, as well as the scanned images of the registers are all records." | "The ordonnances and inscriptions created within the Alsace-Moselle computerized land registry fulfill all the requirements of a record, and may be considered as such. Strict procedural and documentary controls ensure that records are reliable, and there are procedural and technological controls in place to ensure authenticity of the records over time" (Douglas 2006). |
| CS19 | Authenticating Engineering Objects for Digital Preservation (Hawkins 2005) | "The creator considers the generated drawing to be the record of definition, not the model even though any change to the drawing requires first a change to the model followed by a regeneration of the drawing. There are other digital object files that are generated from the solid model, such as tool path files and inspection path files which are considered | "According to the above analysis, the digital entities comprising the "bill of materials structure," 1) and 2), as set aside during the business activities of the originating research partner, as well as the digital entities comprising the test records 1) |

| | | | |
|---|---|---|---|
| | | records. The problem may be cultural more than technological. Engineers and craftsmen still prefer to see a drawing spread out versus looking at a tiny screen." | through 5) generated and evaluated during the engineering experiment, have met all the requirements of a record as defined by InterPARES 1" (Hawkins 2006). |
| CS24 | City of Vancouver Geographic Information System (VanMap) (McLellan 2005) | There was no consensus among respondents of what is considered to be a record. | "According to the above analysis, VanMap cannot at this time be considered a record. Instead, VanMap is a potential record that exists in a perpetually "live" state. In order to be considered a record, VanMap will have to be artificially closed at determined times, and set aside as a record. In addition, written procedures for each activity that involves consulting VanMap will have to be developed and implemented" (Douglas 2006). |
| CS26 | Most Satellite Mission: Preservation of Space Telescope Data (Ballaux 2005) | "They consider the sds and FITS files as the most important files that are created. The raw data (sds) files contain original data; the FITS files are the first instantiation in which the data are represented in a comprehensible way. However, the researchers indicated that it is possible to recreate all files (with the exception of the raw data files; therefore, no raw data files are deleted)." | "the sds raw data colleted by the satellite and the various data products generated using the raw data are all records of the MOST Project" (Xie 2005). |