From Data to Records: Preserving the Geographic Information System of the City of Vancouver

Glenn Dingwall, Richard Marciano, Reagan Moore, and Evelyn Peters McLellan

RÉSUMÉ VanMap est un système d'information géographique utilisé par le gouvernement municipal de la ville de Vancouver, en Colombie-Britannique. Il permet au personnel municipal en quête de données sur la ville, d'accéder à une grande variété d'information spatiale en constante évolution, puisée de sources internes et externes, qui servent à appuyer les procédés administratifs officiels et non-officiels. Dans la première partie de cet article, les auteurs examinent des questions relatives à la préservation à long-terme de VanMap, questions soulevées lors de l'étude de cas faite dans le cadre du projet InterPARES 2 en 2004-2005. Ils examinent VanMap dans le contexte des concepts élaborés par InterPARES sur les documents dans des milieux dynamiques et interactifs, puis ils se penchent sur les questions qui en découlent, à savoir comment établir les éléments essentiels du système qui doivent être préservés, et comment lier le système préservé aux activités qu'il supportait. La deuxième partie de l'article explore comment la technologie Storage Resource Broker, développée au San Diego Supercomputer Center, pourrait servir comme composante d'une stratégie de préservation complète. Cette technologie pourrait appuyer le maintien de l'intégrité et de l'authenticité du système préservé, et contribuer à l'indépendance de son infrastructure grâce à des caractéristiques comme l'espace de nommage fixe.

ABSTRACT VanMap is a geographic information system used by the municipal government of the City of Vancouver, British Columbia. It provides city staff access to a wide variety of continuously changing spatial information about the city, drawn from internal and external sources, which is used in support of both formal and informal business processes. In the first part of the article, the authors address issues related to the long-term preservation of VanMap encountered in the course of the InterPARES 2 case study conducted in 2004–2005. VanMap is examined within the context of InterPARES concepts of the record in dynamic and interactive environments, and subsequent implications in terms of establishing the essential elements of the system that must be preserved and how to relate the preserved system to the activities it supported are discussed. The second part of the article explores how the Storage Resource Broker data-grid technology developed at the San Diego Supercomputer Center might be used as a component in the overall preservation strategy. This technology would support maintaining the integrity and authenticity of the preserved system, and contribute to its infrastructure independence through features such as persistent namespaces.

ARCHIVARIA 64 (Fall 2007): 181–198

The city of Vancouver is a municipality with a population of just under 600,000[1] located on the west coast of the province of British Columbia, Canada. The city is governed by a council consisting of a mayor and ten councillors, as well as an elected Board of Parks and Recreation, and other elected and appointed agencies. Most local government functions are carried out by a centralized civic bureaucracy under the administrative control of the city manager, who reports directly to council. Under provincial statutes, the municipal government is enabled and/or required to provide services related to the following: physical infrastructure and environment; emergency services; regulation of businesses; regulation of property development; waste disposal; animal control; water distribution; parks and recreational facilities and programs; arts and cultural programs, including library, archives, and civic theatre management; and some social planning and social housing programs. In 1999 the Information Technology Department within the civic administration launched VanMap, a corporate Geographic Information System (GIS) designed to integrate information from various departments and display this information in a graphic, interactive form for use by city staff.[2] VanMap was accepted as an InterPARES 2 case study[3] in February 2004 and subsequently became the subject of a collaborative research effort between the City of Vancouver, InterPARES 2, and the San Diego Supercomputer Center, the purpose of which was to propose and test possible preservation strategies for the system.[4]

A GIS is designed to render as maps geographically-referenced data (translated into lines, points, symbols, and shapes), vector-based graphics, raster images, and textual information. Typically, the data are presented to the end

1 The 2006 population estimate is 587,891. BC Stats, "BC Municipal Population Estimates, 1996–2006, Sorted by Name," available at http://www.bcstats.gov.bc.ca/data/pop/pop/mun/Mun9606a.asp (accessed 5 September 2007).

2 A limited version of VanMap has been made available to the public on the Web since 2002, at http://www.vancouver.ca/vanmap/ (accessed 5 September 2007). However, throughout this article, "VanMap" refers to the internal version that is used by city staff.

3 InterPARES (International Research on Permanent Authentic Records in Electronic Systems) was a UBC-led research project running from 1999 through 2006. InterPARES 2 (2002–2006) investigated issues of authenticity, reliability, and accuracy of electronic 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 5 September 2007).

4 Some relatively limited research has been conducted on the preservation of GIS's, resulting in practical guidelines for ingest of GIS data, metadata requirements, and the selection of preservation formats. See, for example, Jo Clarke and Jenny Mitcham, "Preservation Handbook: Geographical Information Systems," *Arts and Humanities Data Service* (June 2005), http://ahds.ac.uk/preservation/gis-preservation-handbook.pdf (accessed 5 September 2007); and Centre for International Earth Science Information Network (CIESIN), Columbia University, "Guide to Managing Geospatial Electronic Records," (June 2005) http://www.ciesin.columbia.edu/ger/GuideToManagingGERv1Final.pdf (accessed 6

user as layers, which may be turned on or off depending on the type of map desired. The end user selects the layers and is able to zoom in and out of various geographic areas presented in map form. In the case of VanMap, the geographic region represented is the entire city, and city staff are able to view the geographic area as a whole or zoom in to view individual streets and neighbourhoods in considerable detail. As of January 2007, 260 layers were available for viewing. These layers include city boundaries; street and property locations; contour lines and other physical features; parks and recreational facilities; water mains; sewer lines; gas lines and other utilities information; traffic light and sign locations; fire hydrant locations; zoning districts; bicycle routes; neighbourhood and administrative boundaries; business licenses; property tax information (presented in tabular form when the user clicks on individual properties); and numerous other types of information related to the city's infrastructure and built environment. Also available are comprehensive orthographic photographs[5] from 1999, 2002, 2004, 2005, and 2006. All the layers, including the photographs, can be viewed superimposed on other layers to allow for the visual correlation of disparate types of data.

VanMap is made available to city staff through a Web-browser interface, which accesses software that aggregates data and digital objects from a variety of sources to form interactive maps. The main data source is an Oracle-Spatial database, which contains information entered directly by staff in various contributing departments, or extracted or copied to it from other sources. Some data are drawn from other systems, such as smaller departmental databases, or from files such as HTML pages, graphics files, and the orthographic photographs. The data themselves are either created by city staff in the process of conducting civic business or acquired from external sources such as utility companies, the provincial government, or crown corporations. In both cases, the data are supplied to VanMap by departments in order to allow them to view their own data in map form, correlated with data from other departments as needed. The system thus provides a centralized repository of

September 2007). Major research projects currently underway in this field include the National Geospatial Digital Archive led by Stanford University and the University of Southern California at Santa Barbara, http://ngda.org/ (accessed 6 September 2007); the North Carolina Geospatial Data Archiving Project, http://www.lib.ncsu.edu/ncgdap/index.html (accessed 6 September 2007); the eLegacy Project, a collaboration between the San Diego Super Computer Center, the State of California Archives, and California Environmental Resources Evaluations Systems (CERES); and Maine State Archives GeoArchives, http://www.maine.gov/sos/arc/GeoArchives/geoarch.html (accessed 6 September 2007).

5 An orthographic photograph (or orthophoto) is an aerial photograph that has been processed in such a way as to eliminate image displacement due to camera tilt and terrain relief, so that it represents every object as if viewed directly from above.

information that can be accessed quickly and can be easily customized to meet a wide variety of end-user needs. Because the system is both comprehensive and accessible, VanMap has become a vital resource for staff as they provide services to the public, enforce city bylaws, and make decisions relating to civic programs and services.

VanMap has grown considerably since it was first launched, with new layers being added every few months. The addition of new layers to VanMap is an informal process. Typically, departments that create or receive data related to one of their business activities, recognizing that there is a spatial component to that data, approach the VanMap team, and ask that the data be added to VanMap as a new layer. The request is made either because departmental staff wish to view this data in context with other spatial information already available through VanMap for their own purposes, or because they feel that staff in other departments may find the new data useful for their own work. VanMap data are incorporated into the system both because there is a specific or essential business requirement that the data be accessible through a GIS, and because it is recognized that geospatial representation of the data in context with a wealth of other data accumulated by the city allows the information to be used in new and often unanticipated ways by all city staff and contributes to the enhanced delivery of public services.

At its most basic level, VanMap provides evidence of transactions (such as the installation of a fire hydrant or the issuance of a business license) which might be better documented elsewhere. However, the system also aggregates and presents information in ways that facilitate new activities and transactions. For example, VanMap can participate in the action of planning the location of a new park or community centre by presenting, in graphic form, information about a neighbourhood's physical layout, traffic patterns, rate of growth, types of businesses, proximity to bicycle routes, and so forth. VanMap also provides functionalities which may not be available in other systems; an example is view cones, which allow a user to calculate acceptable building height based on the extent to which a building might obscure the view of the north-shore mountains, an important factor in the approval of development permits. The system thus provides support for and (to a lesser extent) evidence of specific, localized activities such as the installation of a traffic light, while also being a resource for ongoing activities such as city planning and large-scale infrastructure development.

Surprisingly, however, VanMap offers limited ability to track change over time, for the simple reason that, with some exceptions, when new data are added to the system the old data are overwritten or deleted. For some types of data this process can result in frequent, ongoing change. For example, the city undergoes numerous changes in zoning on a regular basis to accommodate new development; the new or altered zoning districts are entered into VanMap and the old zoning districts disappear. VanMap is thus a "live" system, provid-

ing detailed information on the city of Vancouver at the present moment at any given time, but not saving the information for future use. Moreover, city staff do not typically save the views of VanMap they use to conduct a transaction or make a decision.[6] Thus a map that is created as a result of the interaction between user input and the data available at a specific point in time exists for only a few moments and then vanishes.

This ephemerality poses interesting conceptual and technical challenges to any preservation efforts directed at VanMap. The main conceptual difficulty lies in determining exactly what constitutes the record or records in this system, since records are what archivists seek to preserve. InterPARES defines *record* as "[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," and *document*, in part, as "[an] indivisible unit of information constituted by a message affixed to a medium. … A document has fixed form and stable content."[7] It can reasonably be argued that VanMap is an indivisible unit of information because it is created and meant to be used in its aggregated form: it is the totality of the information and the way in which that information is correlated that constitute VanMap, rather than any individual component such as the data layers, graphics, and so forth. The map views created by the user could also be considered indivisible units, since they are complete in themselves as they are created and used (however briefly) to conduct civic business. VanMap also meets the definition of record to the extent that the information it contains is created in the course of conducting the business of the city or is used to facilitate the conduct of this business. In other words, VanMap is both an instrument and a by-product of the practical activities of its creator, the government of the city of Vancouver. However, VanMap clearly lacks key characteristics that would allow it to meet the InterPARES definition of a record: since data are regularly overwritten and since city staff do not save the map views they use to conduct their activities, there is no stable content and nothing is set aside for action or reference. VanMap as a record is both *made* at each update of data and *received* at each use, but never *set aside* during the course of conducting the activity in which it participates. Thus, it never becomes a record in the context of any given business transaction and because of this is not preservable as evidence of the

6 For some types of business processes staff may print out views of maps, or cut and paste them electronically into other documents. They may also save maps to network drives on an ad hoc basis for further action or reference. However, there are few formalized procedures for doing so and there is no systematic capture into an electronic records management system or other organized system that manages electronic records along with their metadata.

7 The InterPARES 2 Project Glossary, http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=10212&CFTOKEN=78755993 (accessed 6 September 2007).

city's activities. The preservation challenge for VanMap, therefore, is to develop an environment in which the content can be set aside, in other words, to add the fixity that the system requires for it to produce records.

In the course of the research conducted by city archivists, InterPARES 2, and the San Diego Supercomputer Center, two possible approaches were proposed. The first was to consider each viewing of VanMap by a staff member in the course of an activity to result in the creation of a discrete record, consisting of the various digital components that comprised the user's view when they consulted the system. It was proposed that a record of the user's interaction could be created by introducing a means of setting aside the digital components – preserving only the information necessary to recreate the specific view of the system, together with metadata that would create an archival bond by linking the record to the action in which it participated. This approach has a number of inherent difficulties. From a practical standpoint, saving the map views into a record-keeping system – that is, saving the map view and adding the metadata necessary to connect it with the specific transaction for which it was created – would be a significant departure from the way city staff currently conduct their business and would likely place an unacceptable procedural burden on them.[8] This would be particularly true for transactions where multiple views of the same area at different scales and with different layers were used to complete the transaction.[9] From a theoretical standpoint, it would be questionable for the preserver to demand that a record be created in the process of conducting an activity strictly for purposes of preservation rather than to facilitate the ability to carry out the activity. Such a record would not be a natural by-product of a transaction, since the creator does not need to create the record to complete the transaction. Moreover, this approach would preserve the set of individual views created by various users in the course of their work, but would not preserve VanMap, which was created and intended to be used in its entirety. To return to the definition of the record, such an approach would preserve the way VanMap was *received* in the course of business but not the way it was *made*.

In order to preserve VanMap in its entirety, it is tempting to consider

8 Some metadata could of course be generated automatically by the system when individual map views were saved, including the time of capture, information about the individual saving the record, and the types of data being viewed. However, the staff member would be required to add certain metadata manually, such as the reason for capture or the type of activity being conducted, and the relationship of the map view to other records (i.e., paper records or electronic records in other systems) produced during the process of completing the activity.

9 Some processes that VanMap supports, particularly those related to planning and development, can unfold over a very long period of time. In these cases it would be impractical to capture every map view consulted and to attempt to record the specific aspect of the process to which that particular view pertained.

saving all of its data at regular intervals, such as once a week or even once a day. While conceptually simple, this "snapshot" approach would be problematic because the various layers in VanMap are updated with different frequencies. Some layers, such as city boundaries and orthophotos, are never changed.[10] Some layers are updated periodically by accumulating changes in the data and then making a batch update to the layer; for example, changes to city-owned properties are updated in the system once a month. On the other hand, the development permit layer is updated daily to show all of the changes from the originating departmental database made the previous day, and the public streets network layer shows changes immediately after they are made in the Oracle Spatial database that contains the street-network data. Thus, saving VanMap at regularly-scheduled intervals would repeatedly save many layers that have never been updated or were not updated recently, and would miss some updates for layers that are changed frequently.

The researchers' second proposal, therefore, was to save each layer every time it is updated. When a user wished to see what VanMap looked like on a given date (and, potentially, even at a given time), for each layer that the user wanted to view, the system would retrieve the last version of the layer captured up to that point in time. Unlike the first proposal, this would not create a record of VanMap's participation in a transaction, but would create a record of VanMap by allowing it to be reproduced in its entirety as it was made with each update to the data. This approach would have two major advantages over the first proposal: it would not require staff to alter their business processes to produce records for the sole purpose of preserving them, and it would capture VanMap in its entirety, as it was created and as it was intended to be used.

The most serious objection to this approach is that, because the map views the user creates would not be saved, it would not allow for the connection to be made between VanMap and the various transactions in which VanMap participates. In other words, to return once again to our definition of record, it would preserve VanMap as it was *made* but not as it was *received*. This problem would need to be addressed through improved business-process documentation for the city staff who use VanMap to conduct their activities. This is the solution suggested by Luciana Duranti and Ken Thibodeau in "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES." According to Duranti and Thibodeau, this documentation would consist of both a detailed description of the business process and how VanMap participates in it, and instructions on how to create the

10 Each new set of orthographic photographs is added as a separate layer corresponding only to the year in which they were taken.

records supporting and documenting this process. This would allow VanMap to be set aside in the context of the various activities in which it participates, since "the description would reveal the archival bond between the records of each business process and VanMap and the instructions would reveal the specific relationship between each process and the data which supported it."[11] For example, if VanMap were used to determine whether or not to issue a business license or a building permit, the process by which the system would be used, and the types of records that would result, would be documented in a set of procedures staff would be advised or required to follow. Such documentation already exists for a number of business processes but would need to be added or enhanced for others.[12] Of course, while business-process documentation could be developed for certain regulated, formalized processes such as the issuance of a permit or the enforcement of a bylaw, more ad hoc uses, such as informing a citizen about the location of a traffic circle or checking the types of trees in a park, could not be proceduralized in this way. However, it is the regulated, highly formalized transactions which have the most stringent requirements for records creation; ad hoc information gathering, although it may ultimately support such regulated processes as planning and development, does not demand this in the digital environment any more than it does in a paper records environment, and for such uses it would be enough to know what information was available in VanMap at a given time without knowing exactly which layers were viewed, at what scale, and so forth.

The detailed analysis that the development of such business-process documentation would require would yield another beneficial result. For some layers it may be the case that although the data change frequently, the nature of the changes or the impact those changes have on the processes they support would be such that the layer would not need to be preserved after every change to the data. An understanding of the types of information required for the highly regulated, formalized processes with high requirements for records creation and preservation would dictate that certain layers must be saved each and every time they are changed. If it became clear that other types of information were used only for ad hoc information gathering not directly related to specific, formal transactions, it would be reasonable to make the appraisal decision that the layers containing this information could be saved only when certain types of updates were made to it or at regularly scheduled intervals.

11 Luciana Duranti and Ken Thibodeau, "The Concept of Record in Interactive, Experiential and Dynamic Environments: The View of InterPARES," *Archival Science*, vol. 6, no. 1 (March 2006), p. 65.

12 Creating mailing lists using the VanMap notification application (which allows users to create mailing lists generated by selecting a geographic area in VanMap) is an example of a business process facilitated by VanMap that is highly controlled.

Moreover, if the time frame involved in a decision-making process was substantially greater than the frequency of changes to the layers supporting that process, the numerous small changes that occurred in VanMap during that process might be considered inconsequential. This might be the case for ongoing, continuous activities such as long-range planning and development. Each layer would need to be appraised within the context of the business processes it supports in order to determine how frequently and under what circumstances to save the changed versions.

Determining what should be saved and set aside as the record is the first step in preserving records in the complex, dynamic environment of the VanMap GIS. The next phase is to consider the kinds of technological solutions required to implement what we have so far laid out in purely conceptual terms. For this, we turn to the data grid technology developed by the San Diego Supercomputer Center. Data grids are software systems that manage distributed records. The records may be located on any type of storage system, distributed across multiple sites, and even distributed across multiple administrative domains (in the case of VanMap, different city departments). Data grids provide uniform access mechanisms across the heterogeneous storage systems, in part through the use of name spaces, which are global, permanent identifiers for attributes independent of the creator's record-keeping system. Name spaces may be used to identify the records or digital objects (the logical file name), storage resources, and users (the distinguished user name space). The logical file name space makes it possible to organize logical collections of records (i.e., collections of digital objects aggregated on the basis of some descriptive attribute) independently of where the records are stored. Descriptive metadata can be associated with each record, such as the name of the organizational unit that provided the data, the date the record was created, the person who initiated the record creation, and any other desired attribute. The logical name space for storage resources simplifies the maintenance of the records by supporting collective operations across multiple storage systems. Examples are load levelling, the distribution of records uniformly across multiple storage systems; replication, the automated creation of copies of the records on multiple storage systems (reducing the risk of data loss and minimizing the impact if a storage vendor goes out of business); fault tolerance, the ability to store records on the storage systems that are available; and reliable access, the ability to switch to an alternate storage system if the initial request is to a storage system that is off-line for maintenance. The logical name space for users makes it possible to authenticate access to the preservation system independently of the administrative domain managing the storage systems, a capability referred to as "trust virtualization." Access controls are implemented as constraints between the logical file name space and the logical user name space, with the controls being used to restrict the ability to make changes to designated archivists.

Data grids have a major role to play in digital-records preservation. Trust virtualization permits the authentication of users to verify their identities and check what operations they are authorized to perform before permitting access to the data at the remote site. In effect, the data grid acts as the surrogate for the archivist, applying the preservation controls needed to ensure authenticity. Additionally, because persistent names are managed for both the records and the archivists, audit trails can be managed that record all preservation operations that are performed. The data grid ensures that as the records are moved between storage systems under data-grid control, the access controls do not change. This enables the tracking of the chain of custody: the identification of all institutions and archivists that have stored or managed the records. Most significantly, data grids are able to manage the records, metadata, access, and storage through successive migrations of the underlying data or through the addition of new technology to the preservation environment: when new technology is added to the preservation environment, the data grid can access both the old and the new environments simultaneously, and thus can be used to migrate data from the old technology to the new technology, a capability referred to as "infrastructure independence."

The de facto standard for data-grid technology is the Storage Resource Broker (SRB), developed at the San Diego Supercomputer Center.[13] The SRB is used as a generic, distributed data-management infrastructure to support collection building (from real-time sensor data streams),[14] data sharing (data grids), data publication (digital libraries), and data preservation (persistent archives, based on the infrastructure independence capability described above).[15] The SRB technology is used on an international scale, with shared collections created across storage resources in multiple countries. An example is the US National Optical Astronomy Observatory (NOAO), which uses the SRB technology to manage data taken by telescopes in Chile and archived on storage resources in the United States. Large-scale data-management systems rely upon the concept of federation: multiple independent data grids are creat-

13 SRB software is currently available on a limited basis to educational and US government agencies. See http://www.sdsc.edu/srb/index.php/Is_SRB_Open_Source (accessed 6 September 2007). A commercial version is also available.
See http://www.nirvanastorage.com/index.php?module=htmlpages&func=display&pid=1 (accessed 6 September 2007).

14 Collection building refers to the organization of observational data and the association of descriptive metadata that can be queried. In the real-time sensor community, over 4,000 data streams are organized into a single collection that can be browsed to discover relevant data streams.

15 For more on persistent archives, see Reagan Moore, "Building Preservation Environments with Data Grid Technology," *The American Archivist*, vol. 69, no. 1 (Spring/Summer 2006), pp. 139–58.

ed, each with their own set of logical name spaces and their own metadata catalogue. Federation is the controlled sharing of name spaces between the independent data grids. It is possible to control the cross registration of the logical user name space between two data grids, copy files between data grids, copy metadata between data grids, and even control whether storage resources will be shared between data grids. The NOAO example actually federates five independent data grids to manage the distribution of data between Chile and the US, with data being pulled from data grids in Chile onto data grids in the US.

Federations are also used to build "deep archives," preservation environments that ensure that only the archivist is able to access the preserved data, minimizing the chance that an unauthorized user might change an archived record. Deep archives are constructed by the federation of three data grids: a publicly-accessible data grid; an intermediate staging data grid that is only accessible by an archivist; and the deep archives data grid. The name of the archivist controlling the staging data grid is registered into the public data grid, enabling the archivist to pull data from the public data grid into the staging one. Similarly, the name of the archivist controlling the deep archives is registered into the staging data grid to permit them to pull data into the deep archives. The result is that the identity of the deep-archives archivist and the location of the deep archives are unknown to users of the public data grid. Through the use of firewalls and virtual private networks (communications networks that mimic restricted, private networks but which are carried over a public networking infrastructure such as the Internet), all access to the deep archives can be controlled, improving the ability to securely manage the records.

In the case of VanMap, data-grid software could be inserted between the data storage systems and the access applications used to identify and retrieve the records, and each saved layer within the GIS could be independently registered in the data grid (this would happen automatically when the layers were saved). The layers could then be organized into a logical collection hierarchy based on the dates on which they were saved. Data grids support logical soft links, which are pointers to digital objects that have been captured in other logical collections; this means that a user could enter a query for a specific date into a search engine, and the data grid would retrieve all layers that had been saved on that date, plus, using the logical soft links, the most recently changed versions of other layers that had not been saved on the exact date queried. The result would be a composite view of VanMap, which, in terms of content, would be similar or identical to the VanMap that a city staff member would have seen on that date. This would allow for the recreation of VanMap as it was used to support a specific decision or transaction. Of course, other types of queries could be built as well. For example, a user might be able to select a desired layer and compare it as it was saved on

different occasions. This would be accomplished by retrieving the different versions and building a presentation that identifies differences between the layers. In the same way, different "snapshots" of the system, consisting of comparisons of multiple layers selected for different dates, could be compared. These capabilities would give insight into the types of information that were available to city staff as they conducted ongoing activities such as long-range planning and development, for which viewing the layers only on specific dates might not be adequate to understand the process as it unfolded over time.

To test the ability of data-grid software to manage VanMap across a migration, a subset of layers was captured on two different dates for storage in an SRB based at the San Diego Supercomputer Center.[16] The VanMap GIS has two principal components: a file that defines how the layers are composed and properties for each layer (such as line colour and the symbols used to represent data); and a file that contains the data presented in the layer. Each of these files was stored in the SRB data grid with temporal tags, and a date-based query on the system successfully evaluated the closest previous version of the layers, retrieved the GIS files, and loaded the information into an open-source GIS.[17]

The test thus simulated what might result from a typical migration of the data from one software environment to another. It is assumed that the technol-

16 The files obtained were meant to be a representative sample of six common types of geospatial data found in the system: (1) Access-database files (.mdb); (2) Oracle-Spatial database exports (.exp); (3) Autodesk MapGuide native SDF files (.sdf); (4) Coldfusion reports (.cfm); (5) Autodesk MapGuide Map Window Files, the project files that link all the layers together and define a particular map (.mwf and .mwx); and (6) raster-image files (.ecw). The actual data represented the following layers: (1) parcel-lot polygons with tax attributes; (2) orthophotos; (3) zoning districts; (4) city-owned properties; (5) city projects; (6) traffic-related parking-meter rates; (7) traffic-related traffic counts; (8) public streets; and (9) webcam images.

17 San Diego Supercomputer's test-bed environment included a full version of Autodesk MapGuide 6.5 (to mirror the city of Vancouver's own system) and an ArcIMS GIS server (to represent a completely different commercial GIS-software product). The layers obtained from the city were loaded into the Autodesk MapGuide instance, to simulate the original environment, set aside in the data grid and "decorated" with additional temporal attributes, and when queried in the SRB system, retrieved and loaded on the fly into the ArcIMS instance. This was done by creating the equivalent of an Autodesk MapGuide .mwx file (.axl in the case of ArcIMS) on the fly and launching a map service with those retrieved layers and serving them on the Web to a standard web browser. At that point the user can interact with the recreated historical map from the browser, thus simulating the experience of viewing a legacy VanMap. The researchers also experimented with geospatial formats by converting the original formats into alternative forms (this was done using the FME software (feature manipulation engine) from Safe software, a British Columbia company specializing in solutions to translate and transform spatial data formats).

ogy used to create and view information in VanMap will change over time, requiring that the VanMap data be migrated into a preservation format. However, any migration raises the issue of authenticity. The preserved version of VanMap can be said to be authentic if it communicates the message originally intended by the creator of the original. In order to be able to declare the preserved VanMap authentic, we need to establish that the information content of the preserved layers has not changed, and that the elements of the documentary form necessary for the authentic communication of the content are present in the preserved record. Furthermore, we need to show that the spatial correlation of the layers, and the interactions between them, have been retained, and that the manner in which the application used to view the layers is equivalent to the original presentation.

Documentary form is defined as "the rules of representation according to which the content of a record, its administrative and documentary context, and its authority are communicated."[18] Documentary form may include such extrinsic elements[19] as colour or line thickness, if these are necessary to convey meaning. In the case of VanMap, an appraisal must be conducted on each layer to establish what elements of documentary form, if any, are essential to the proper communication of the information content of the layer in question. Consider the example of elevation contour lines. The layer that shows one-metre contour lines is colour coded, using blue lines for lower elevations and progressing through green, yellow, orange, and red lines as elevations increase. The importance of this aspect of the extrinsic form to the communication of elevation information would need to be considered when making decisions about what to preserve. Would the preserved version be considered authentic if only the spatial information were preserved (i.e., the positions of the contour lines but not the original colouration of the lines)? If it were determined that different line colours are a critical component of the record's documentary form, would it be sufficient that the lines be coloured differently from each other or would the original colour scheme need to be preserved as well? In making these decisions, we must also consider how the elements associated with each layer interact with each other. Taking the example of contour lines again, there are layers that show elevation information in one-m, two-m, and ten-m intervals. The one-m and two-m layers use contour lines that are one pixel wide, while the ten-m layer uses lines that are two pixels wide. When a layer is viewed alone, the width of the lines is largely irrelevant to communicating elevation information to the user. However, when

18 InterPARES 2 Project "Glossary."

19 An extrinsic element is "An element of a record that constitutes its external appearance." Ibid.

a user views the one-m and ten-m layers simultaneously, the line thickness becomes relevant as it helps to distinguish the many one-m contour lines from the fewer ten-m contour lines.

With regard to the spatial correlation between the layers, the behaviour of the system when different layers are superpositioned is one of the most important features of a GIS. If the information in a layer is migrated from its original format to a preservation format, care must be taken that the migrated layers are properly aligned when viewed in the preservation environment. The importance of this can vary from layer to layer, depending on how precisely the spatial information needs to be displayed in order to adequately communicate its content. This must be considered not only within the context of the layer itself, but also in the context of how the accuracy and precision of the spatial information in a layer affects the interpretation of other layers. Finally, any application used to view the layers must be capable of presenting the layers in the same way that they were originally viewed by VanMap users. If the application used to view the preserved layers limited the number of layers that could be viewed simultaneously (to fewer than the original application), or if the application improperly applied the scales at which different layers "drop in" or "drop out" (become viewable or not viewable), the recreated record could not be said to be authentic.

The selection of the elements that must be preserved (kept invariant) thus needs to be defined for each layer. Information about these elements can be recorded in a template and stored along with each layer in the preservation environment; when the layer is displayed, the template is used to control the presentation properties. However, information about the behaviours associated with the interactive maps must also be preserved. In practice, as technology evolves, the interaction mechanisms become more sophisticated and may require associated display parameters that were not managed for old forms of the technology. In this case, users may want to apply new behaviours on data from prior GIS environments, but from the standpoint of preserving VanMap as evidence of the activities of the city of Vancouver, it is important to fully understand the behaviours that were available to city staff at the time they used the system. A harder problem to manage is when behaviours are deprecated and no longer supported by newer technology.[20] One approach to handling this problem would be to provide as a minimal set of behaviours those present within the current VanMap system, identifying the software specifications that support those behaviours. As later versions of the software

20 An example of this problem occurred with the replacement of VRML (Virtual Reality Modeling Language) version 1 standard with the VRML version 2 standard: little-used features were removed in the process.

become available, any changes in their specifications would be tracked in order to ensure continued support for the minimal behaviour set. Deprecated behaviours would be retained and replicated in the preservation environment, ensuring that the preservation environment provided, at a minimum, the behaviours used in the current system.

We have noted that some of the VanMap layers change much more slowly than others and have recommended as a preservation strategy saving layers each time they are updated. The layers may be saved within the current VanMap system or extracted to another system, which would contain only the saved layers; either of these is a viable approach.[21] The current VanMap environment manages about one terabyte of data, with most of this in the orthophotographic images of the city, and one million files. If layers in the VanMap GIS are saved each time they are updated, the volume of data added per year, assuming about fifty layers are archived each week, is approximately 2.5 gigabytes and 25,000 files. Thus, preserving updated layers would cause a minimal impact on the management of the preservation environment. Current preservation environments based on the SRB manage fifty million files and hundreds of terabytes of data.[22] Without data grid technology, it would be necessary for a central repository to be created, into which each department would deposit standard-database dumps, making it possible to record all changes to government information. A VanMap interface to the central repository would then be used to extract information across all layers for any point in time. This would require a high degree of coordination, a standard-database infrastructure, and the implementation of a central repository. However, with a data grid it becomes possible to register data from multiple independent repositories into a logical collection that is accessed by the VanMap presentation system. If an interface is constructed between VanMap and the data grid, it will then be possible to handle multiple types of data repositories. Data can be retrieved from the current systems or the preservation environment to

21 The ability to save and (to some extent) manage historical data layers may become a common feature of GIS products in the future. It is interesting to note that ESRI's ArcGIS, a widely-used GIS application, has introduced version control capability in its latest release.

22 An alternative preservation strategy considered by the researchers was to preserve daily snapshots of the VanMap GIS layers, with the exception of the orthophotos, which are the largest files and which are never updated. The exclusion of the orthophotos from the daily snapshots would make the amount of data being added to the preservation environment manageable: assuming about one hundred layers archived each week, the volume of data is approximately five gigabytes and 30,000 files. Backups of the underlying databases would also be captured in order to identify when a particular change occurred. The database dumps would need to be stored along with the saved layers in the data grid and the same VanMap interface could be used to access both sets of data. This strategy is thus technologically feasible, but does result in unchanged layers being saved repeatedly and would miss changes to layers where more than one change took place on the same day.

compose the desired presentation. This represents the most sophisticated form of a preservation environment, with the preserved records themselves distributed across multiple repositories. This is a preferred implementation because it minimizes risk of data loss due to natural disaster (through geographic replication), operational error (through replication across administrative domains), vendor product failure (through replication across multiple vendor products), and malicious users (through replication into deep archives).

By introducing fixed content and stable form to VanMap and managing the data through the use of a data grid, we can create a record-keeping environment for the "live system" that is currently VanMap and provide a means of managing it through changes in technology. However, we also need to add processes that manage assertions about authenticity and integrity. To enhance the ability of data grids to function as preservation environments, the San Diego Supercomputer Center has developed iRODS, the integrated Rule-Oriented Data System, a type of adaptive middleware that allows user communities to create and implement customized preservation practices on selected bodies of records within the data grid.[23] These sets of customized rules are programmed as micro-services, small sets of well-defined procedures and functions that perform specified tasks. For example, a user may decide to implement a procedure to place more stringent access restrictions on a particular set of files in a collection while leaving the access restrictions for the rest of the files unchanged, or to extract certain types of metadata elements from pre-designated file types. iRODS allows users to create preservation rules to ensure the authenticity and integrity of records without having to make any programming changes. These rules can be expressed as a set of management policies that must be followed to ensure the sustainability, governance, authenticity, and integrity of the preservation environment. The results from application of the rules are stored as persistent state information (i.e., preservation attributes or metadata), which can subsequently be queried and evaluated to prove that the management policy was enforced. The types of management policies that can be applied include:

- Periodic checks of the integrity of the records (evaluation of file checksums)
- Periodic checks of the authenticity of the records (evaluation of the preservation metadata associated with each file)
- Replication validation for required number of copies
- Distribution validation for the required number of independent storage systems

23 Version 0.5 of iRODS was released as open-source software in December 2006. See http://irods.sdsc.edu/index.php/Main_Page (accessed 6 September 2007).

- Audit-trail assessment for actions performed upon each record
- Conformance of the system to submission agreements
- Conformance of the system to disposition agreements
- Conformance of standard reports
- Migration strategy
- Transformative migration strategy for changing file formats
- Presentation strategy for formatting data for display.

The Storage Resource Broker has implemented a standard set of operations for the manipulation of remote data files. These operations have been proven over time to provide the necessary data manipulations. The initial micro-services that have been implemented in the iRODS system provide similar capabilities, based on an analysis of the National Archives and Records Administration (NARA) Electronic Records Archives capability requirements, which has identified 174 rules that should be implemented, ranging from control of individual micro-services and comparison of information content based on templates to validation of preservation operations.[24] However, the goal of the iRODS system is to provide the ability to specify community and collection-specific management policies as rules on micro-services, and to support dynamic extensions to these policies over time. In addition to the logical name spaces managed by an SRB (for users, storage resources, and the digital entities or files themselves), iRODS adds those for rules, micro-services, and persistent state information (preservation attributes). The logical name spaces enable the evolution of the data-management infrastructure itself. A user can add new rules, micro-services, and persistent state information to support a new management policy, and enforce the new management policy on top of the prior management policies without having to modify any code. An iRODS system is thus intended to support management of its own evolution, as well as enforcement of management policies for each community and collection. In the case of VanMap, an example of a rule might be the criterion for the retrieval of a consistent set of GIS layers given a specified date. The rule could specify that each layer be retrieved from the closest prior layer snapshot, or it could specify that only layers from a specified archives be used that meet a set of preservation requirements on chain of custody. The latter might be required for material presented in a court of law, for example, while the former may be used by a city staff for reference purposes.

24 See National Archives and Records Administration, "Attachment 2 to Section J, ERA Requirements Document (RD)," (December 2004), http://www.archives.gov/era/about/requirements-amend0001.doc (accessed 6 September 2007).

The initial interest of the City of Vancouver Archives, prior to its participation in InterPARES, was in preserving the information contained in VanMap because of the perceived historical value of that information to researchers. Putting aside any theoretical discussion of what constitutes a record, it was recognized that preserving a sample of the information in VanMap that could be repurposed by future researchers was a worthwhile endeavour. Further investigation into the role VanMap plays in a wide variety of city business processes quickly led to the realization that VanMap was an important record bearing evidential value about those processes, and should be preserved as such. This seemed a problematic endeavour for a number of reasons, not the least among them being the result of the diplomatic analysis conducted as part of the InterPARES case study that suggested that VanMap was not actually a record, due to its lack of fixed form and stable content.

In order to preserve VanMap as a record, it first needs to become a record. This can be accomplished by having VanMap's creator make a decision to save it, and by the formalization of a number of business processes through the articulation of their relationships with VanMap. It is not enough to decide to save VanMap – we must also take action to do so. This is a daunting challenge given the distribution of the data across multiple repositories, and the need to preserve the internal relationships among the layers as well as the external relationships that layers and sets of layers have with the activities in which they participate. Data-grid technology (including both SRB and iRODS) has shown to be a suitable technology for capturing VanMap components as they are regularly generated and set aside, and for managing the preservation of the components across time in a way that enables the reconstruction of an authentic version of VanMap. The salient features of the technology are the ability to support the interfaces needed to integrate distributed data under preservation systems such as GIS environments, and the ability of data grids to enforce the properties required for preservation while supporting indexing and organizational structures managed by digital libraries. We believe it is possible to build a VanMap system that accesses data from distributed repositories and preservation environments and that can support multiple types of access and display for comparing current and historical records. The transformation of VanMap from a system that organizes spatial data, to a system that *is* a spatial record would, through its preservation, facilitate a much wider use for both archives users and for city staff. This is consistent with the fundamental goal of VanMap – to make available as much geospatial information as possible to city staff, so that they can use it in any manner they believe will help to deliver services to the public and make Vancouver a better city.