

**US Army Engineer District, Detroit
DACW35-01-D-0003, D.O. 0005
Task 5**

**Information Management Strategy for the
International Joint Commission Lake Ontario-St. Lawrence River Study
May, 2002**

Submitted To:

US Army Corps of Engineers
Detroit District
Watershed Hydrology Branch
Engineering and Technical Services
USAED P.O. Box 1027
477 Michigan Ave.
Detroit, MI 48226
USA

Submitted By:

Pangaea Information Technologies, Ltd.
14 E. Jackson Blvd., Suite 1325
Chicago, IL 60604
USA



EXECUTIVE SUMMARY

The Common Data Needs Technical Working Group (CDNTWG) of the International Joint Commission's Lake Ontario – St. Lawrence River Study was charged with the development and implementation of an Information Management Strategy (IMS). In response the CDNTWG assembled an IMS Team consisting of professionals either participating in the Study or associated with agencies or organizations in the Study region, that have relevant experience in information technologies. With assistance from a contractor, Pangaea Information Technologies, the IMS team has conducted a comprehensive Needs Assessment (NA) and hosted two workshops to aid in the formulation of the IMS. A summary of the needs assessment process and results, a survey of the strategies and policies adopted by other organizations, the strategy alternatives and options associated with the implementation phase and recommendations thereof, are presented in this document. After the Study Board selects the alternatives and options to be implemented, and Study IM policies to support those alternatives and options are endorsed, the CDNTWG will coordinate the development and implementation of a detailed Information Management Plan.

During the IM assessment and analysis phase, several key Study properties important to the Information Management Strategy were identified. They include:

- Modeling and data processing is widely distributed within the TWGs among contractors, affiliated agencies and TWG members; other Study Participants and regional resource stakeholders are also widely distributed.
- TWGs identified specific geospatial and aspatial datasets (both model inputs and outputs) as sensitive, for reasons related to security, proprietary, and liability issues.
- While communication and information/data transfer via email and FTP presently address Study needs, this short-term solution is anticipated to become insufficient within 6 months due to dramatic increases in data volumes and demand for that data.
- The IJC does not wish to serve in a data stewardship or distributor capacity after the Study ends.
- Making the Study process transparent to the public is essential to the Study's success. This includes making model descriptions and datasets accessible to the public for review and evaluation.
- Providing data discovery, evaluation, and access for Study Participants has much potential to reduce redundancy and allow for more integrated modeling across resource sectors. Providing this functionality to the public is also desirable if data sensitivity and security can be ensured.
- Any IM system developed for the Study needs to be reliable, redundant (backed-up), and secure.

One of the principle functions required of a system designed to support the sharing of information is the ability of potential users to learn of the existence of and significant details about data or information. Popularly known as *data discovery*, such mechanisms employ search procedures which match user defined criteria against information held about the data, known as metadata. *Metadata* is essential for the data discovery process. The degree to which an organization generates metadata for data and information determines the effectiveness of the data discovery mechanism that can be implemented.

Data storage, maintenance and access needs require the coordination and integration of the many responsibilities associated with the system and its data. *Data owners* hold the responsibility for data use and maintenance, and they have the authority to define and manage data access and distribution through the application of a flexible security model. Authorized by the data owner, *data stewards* are familiar with the issues and concerns specific to a data set, and are responsible for its day-to-day maintenance. Consistent maintenance is essential to ensure the currency and quality of data. The integration of these responsibilities with the infrastructure and organizational procedures that support the system ensures reliability and sustainability.

The Internet continues to be the most commonly utilized method of distributing an information management system to a large number of dispersed users. Through consistent and well-designed implementation, an information management system can consist of one or many servers and provide simultaneous access to multiple users in many different locations. In establishing a shared environment from which data and information resources can be utilized, a client/server strategy can promote efficient system administration and access. The organization and design of the system will also need to address *extensibility*, the capacity to implement additional technical functionality after the initial implementation phase. Examples of such additional functionality are web services such as *web mapping services* (WMS) and *web feature services* (WFS) which provide functionality for interactive geospatial data viewing and querying over the Internet. This type of functionality would require the implementation of a system that allows for connectivity to multiple datasets, potentially over multiple systems.

With knowledge gained from a review of the IMS approaches and policies implemented by organizationally- and functionally-similar organizations, and the information about Study properties and needs (presented above), the IMS team synthesized a strategy and several specific approaches for implementation. The alternatives and options were divided into three distinct areas:

- 1) Data Discovery
- 2) Data Storage, Maintenance, Access, and Distribution
- 3) Document Information Management

Data Discovery Alternatives and Options

Alternative 1 - Status-Quo: Currently, data discovery performed in the LOSLR Study is a function of gleaning information from documents detailing the Study organization and work plans and/or by “word-of-mouth.” The currency and completeness of this Status Quo approach is often poor.

Alternative 2 - Tabular List of Data: A second alternative would be to generate a tabular list of all data used or generated by the Study, and ask the respective data owners to add some brief metadata to their entry(s). The list could only be distributed to Study Participants because the non-compliant metadata would not be fit for public consumption. This alternative addresses the immediate need for inter-TWG data awareness in a limited manner, but does nothing to promote transparency and openness of the Study for the public.

Alternative 3 - Metadata Catalogue: A third alternative would be to develop a collection of standard-compliant metadata files, or *metadata catalogue*. This alternative is the first to be fit for public consumption, addressing the need for public involvement and transparency in the Study process and thereby promoting its overall credibility. This alternative requires a Study-wide commitment for the development and coordination of standard-compliant metadata, which requires a formal metadata review process.

Metadata Options:

The following options may be selected to further assist in the metadata development and coordination:

Option 1 - Metadata Review Team: The first of these options is the formation of a Metadata Review Team, which would conduct quality assurance and quality control on metadata as it is generated by the TWGs.

Option 2 - Metadata Coordinator: The second option is the hiring of a Metadata Coordinator, who would coordinate all metadata training, provide assistance in metadata development, ensure completeness of metadata produced, and confirm compliance with FGDC 1998 metadata standards.

Option 3 - Metadata Workshop: The third option is to hold a Metadata Workshop for training all study participants involved in the production of standard-compliant metadata. Workshop training on metadata generation software could provide a jump-start to the metadata creation process, and reduce the time spent by a Metadata Review Team and/or Metadata Coordinator over the course of the Study.

Option 4 - On-line Metadata Development Assistance: The final option would be to design and implement On-line Metadata Development Assistance. This service

would help TWGs that are generating metadata through simple text instructions and easy to understand manuals, and to direct specific questions to an identified metadata expert (e.g., the Metadata Coordinator), who would be required to provide timely assistance.

Alternative 4 - Spatial Data Infrastructure (SDI) Participation: The fourth and final alternative that addresses the need for Data Discovery is the participation in the spatial data infrastructure (SDI) of the United States and Canada. The SDI includes a network of metadata providers that use a standard search protocol to allow access to metadata through a single data discovery portal. Participation in the clearinghouse networks requires FGDC- or ISO-compliant metadata, and a searchable server. Thus, this alternative incorporates the tasks necessary for implementation of the third alternative, i.e., production of the metadata catalogue.

Because participation in the SDI network requires the implementation of a searchable (i.e., Z39.50-compliant) server, the Study would most efficiently utilize resources by submitting metadata to an agency or organization who has already implemented a clearinghouse node server. Examples include the Great Lakes Information Network Data Access (GLINDA) Clearinghouse, established by the Great Lakes Commission (GLC), and the Canadian GeoConnections.

Recommendation of Data Discovery Alternatives and Options

To best address the need for *Data Discovery and Evaluation*, implementation of Alternative 4 - SDI Participation is recommended. In addition to its primary function, positive externalities of this alternative for the Study include becoming part of a developing service provided to the geospatial data community, facilitating the transparency of the Study, and enhancing the overall visibility of the Study through its inclusion in the Global Spatial Data Infrastructure (GSDI). Options 2 – 4 are also recommended: hiring a “Metadata Coordinator”, conducting a “Metadata Workshop”, and providing “Online Metadata Development Assistance”. The primary cost associated with this alternative is related to the creation of metadata and the optional support functions. While some support for the SDI node may be appropriate (requested or required), that additional expense would be minimal. The total estimated cost of implementing the recommended Data Discovery and Evaluation alternative and options is \$73,356.00US in FY2002 and \$176,101.50 thru FY2005.

Data Storage, Maintenance, Access, and Distribution Alternatives and Options

Six alternatives have been identified for addressing the needs for data storage, maintenance, access and distribution. While the implementation of multiple alternatives simultaneously was possible for data discovery, the alternatives here are less compatible. The possible exception to this would be the temporary “implementation” of an “extended status quo” to accommodate the

short-term needs of the Study during the development, testing, and final implementation of a better alternative. Additional options have also been identified that could be implemented with the three more functional alternatives

Alternative 1 – Status-Quo: The current data storage and access scheme implemented for the Study allows users (Study Participants) to store and access data in their local environments. Data transfers to non-local users requires an FTP site, such as the one managed by Canadian Centre for Inland Waters (CCIW), or various media (e.g., CDs, magnetic tapes, etc.). The system for data distribution is largely uncoordinated and fails to facilitate data integrity, security, back-ups or archiving. This system includes no active maintenance functionality for individual datasets: incremental changes to parts of a dataset cannot be made, and only wholesale replacement is possible. Considerations for public accessibility of data and long-term sustainability of data and systems are not addressed under the current strategy. No immediate additional costs are associated with continuing with the Status Quo alternative; however, because the CCIW FTP site was intended as a temporary solution, a decision to continue with this strategy will likely require that additional capacity be added in the near future as the demand for its use increases.

Alternative 2 – Single Repository: The second alternative identified to address the need for a coordinated data storage, maintenance, access and distribution is the implementation of a Single Repository for Study data. The repository would exist as single FTP site to which users can be assigned rights and permissions according to their specific information needs. As a single location for all Study data, the repository would allow for much greater coordination of data distribution, and data integrity, security, back-up and archiving would be facilitated. The repository would be able to accommodate public access to data through providing limited access with read-only permissions or by implementing a webpage with hyperlinks to FTP-downloadable files.

Alternative 3 – Single Data Base Management System (DBMS): The third alternative identified to address the need for a coordinated data management strategy involves the implementation of a Single Data Base Management System (DBMS) for data storage, maintenance, access, and distribution. Establishing a Single DBMS in which data is loaded and stored in a logical structure in a relational database environment would allow for data to be integrated into other systems. It would also accommodate the application of other technologies much more effectively than through using the file structure approach of the previous two examples. The single location will facilitate data integrity, security, back-up and archiving. However, because long-term sustainability is dependant upon the willingness and ability of data owners and stewards to maintain datasets, as with the previous alternatives this one prohibits long-term sustainability by inhibiting regional ownership and stewardship. Policies to provide for appropriate public accessibility would need to be established under the Single DBMS alternative. Similar to the single repository, a flexible data security model and standards for data transfer would need to be implemented.

Alternative 4 – IJC Distributed DBMS: The fourth alternative identified to address the need for a coordinated data management strategy involves the implementation of a data system similar to the single DBMS described above, but divided and managed by the respective national offices of the IJC in Ottawa and Washington DC. A dual system would be developed and maintained in a consistent and interoperable manner so as to support seamless data access across national jurisdictions. By committing to the development and maintenance of a system managing data for the LOSLR Study by national jurisdiction, the IJC would build an information management infrastructure to support the data management needs of the LOSLR Study, and potentially, future studies.

This alternative would require the Study Board’s support to equip the IJC national offices with the necessary hardware, software and expertise required to develop, implement and maintain interoperable geodata management systems. Because this approach requires the development of IM support staff and resources, the cost associated with this dual system is substantially greater than the regionally distributed alternative, which takes advantage of the infrastructure and established knowledge base of other regional organizations. However, while the cost is associated directly with the LOSLR Study’s IM system development, implementation, and maintenance, it could also be considered an investment for future studies and other IJC information management needs.

Alternative 5 – Regionally Distributed DBMS: The fifth alternative identified to address the need for a coordinated data management strategy involves the implementation of a data system similar to the Single DBMS described above, but divided and managed at the regional level. The Regionally Distributed DBMS most effectively addresses the need for regional partners to ensure the longevity of data associated with the Study. As with data owners, regional system maintainers would need to be identified just as data owners would. This data management model is the most flexible and progressive; it is endorsed and actively promoted by leaders in the geospatial IT community from the public and private sectors, as well as NGOs.

A Regionally Distributed DBMS would be developed in a coordinated effort to ensure maximum consistency in system implementation and maintenance. Interoperability standards would be need to be specified to ensure greater integration and connectivity to other systems, and can more easily accommodate other technologies such as geospatial web services. At present, probable candidates as regional components in this distributed set of DBMSs include systems managed by the Great Lakes Commission, Land Information Ontario (LIO, a part of the Ministry of Natural Resources) or Environment Canada – Ontario Region, and Environment Canada –Quebec Region (EC-QR). While all three regionally-distributed DBMSs will have separate administration, consistency must be promoted during development to ensure a common approach to data storage, maintenance, access, and distribution. In addition to addressing seamless *system* development and implementation, *data* held in the systems would need to be clipped to a common boundary and otherwise made seamless in order to facilitate the overall consistency of the Study data. System development for this alternative will require investment exceeding that

necessary for the Single DBMS, in order to accommodate for the additional coordination of effort and system implementation.

Alternative 6 – Technical Work Group (TWG) Distributed DBMS: A final alternative that should be considered to address the need for coordinated geospatial data management involves implementing a DBMS similar to the alternative described above, but with components distributed among TWGs. This approach has several advantages, although these are confined to activities that will take place during the duration of the Study. The TWG Distributed DBMS alternative would place the data and system in relatively close association with the data developers and initial data users. As such, reliable access and control over the data has the potential to increase the overall motivation required for system upkeep during the Study. Moreover, because the system and geodata would be managed by that data’s primary user-group, data currency and integrity should remain up-to-date.

However, because this approach includes datasets that encompass international and provincial boundaries, unlike the “Regionally Distributed DBMS” alternative, securing data owners with the motivation to provide for database maintenance beyond the Study’s terminus could prove problematic. Likewise, the system longevity would be dependent on securing a motivated steward prior to the completion of the Study. Obviously, this alternative would require a maximum allocation of funding – required to implement a large network of distributed systems, one for each individual TWG.

DBMS Options:

The three options for the three preceding DBMS alternatives (alternatives 3-5) include:

Option 1 – Data Viewing Map Making: The development and provision for interactive Data Viewing and Map Making using open source software,

Option 2- Proprietary Internet Mapping Service: The implementation of Proprietary Internet Mapping Services that offers more robust geospatial analysis functionality than that in the first option, and/or

Option 3- Middleware: The implementation of system “middleware” that allows the connection of geospatial applications in certain DBMS environments.

Recommendation for Data Storage, Maintenance, Access, and Distribution

To address the need for Data Storage, Maintenance, Access, and Distribution, implementation of a Alternative 5 - Regionally Distributed System is recommended. The system recommended in Alternative 5 would be distributed among the three political regions (Quebec, Ontario, and New York State) that comprise the Study area. Because the IJC does not wish to serve in a data maintenance capacity beyond the life of the Study, data owners and stewards will need to be

assigned to ensure long-term sustainability of data. Regional agencies have the necessary interest in the datasets and motivation to ensure the data's longevity. Hence, this alternative increases the likelihood that the system and the data will likely remain sustainable in the long-term, and can be recommended because of the existing resources available to the study in the form of regional DBMS's and knowledge bases.

DBMS Options 1 and 3 are recommended: establishment of web-based Data Viewing and Mapping capabilities, and installation of "middleware" to provide for system interoperability and OpenGIS Consortium compliancy for other Open Web Services (OWS). Total estimated cost of implementing the recommended alternative and options is \$166,675US in FY2002 and \$312,175 thru FY2005.

Policy essential in the implementation of the recommended alternative and options are:

- All primary Study participants (e.g., Study Board, PIAG, and TWG members) should be given access to all data and information utilized and/or produced by the Study, with the exception of data and information having special security, liability, privacy, licensing, or proprietary concerns.
- All other interested parties should be given access to any data and information which is considered new or having value added to it by activities of the Study, with the exception of data and information having special security, liability, privacy, licensing, or proprietary concerns.
- "New data or information" could be defined as that which did not exist prior to Study activities and was generated from primary data collection procedures as a direct result of Study activities, i.e., model output or results.
- "Value-added data and information" could be defined as that which has been significantly improved as a result of Study activities in either its content or usability.
- Data owners, and especially data steward, should be identified as early as possible prior to the end of the Study.

Document Information Management Alternatives and Options

Without question, other flows of information will be necessary for the Study to be successful. In particular, it is likely that administrative and document management tools will become increasingly desirable as the Study progresses. However, without having developed a Communications Strategy or specific policies for internal reporting procedures or functions, specific recommendations are difficult. Given this lack of information it would be prudent to err towards a more robust document management system that is scalable and possesses the capacity for extensibility. Designing and implementing a system that will not meet changing or currently unforeseen critical Study needs could prove very costly (and wasteful) in the long run. Designing a system that is extensible and scalable provides insurance against this.

Having said this, the following document management system components and functions are tentatively recommended:

- 1) Commercial Off The Shelf (COTS) software for web-based document and other information management, such as Xerox's *DocuShare* (see Section 7.3.2.1.1). This higher-end web-based document management system could prove extremely useful in meeting internal Study IM needs.
- 2) A web-site with documents and other information presented in a hierarchical structure. This is simply a recommendation for the organization of the existing web-site. Basic HTML text search functionality should be provided.
- 3) A web-enabled Shared Vision model. The IMS team views this model as having the potential for more than an excellent decision support tool. Its structure allows the integration of all essential information (i.e., links to model descriptions, model inputs, etc) that could facilitate evaluation of the Study *and* support the recommended data discovery, evaluation and access schema.

System Component Integration

The primary system components, as recommended above, are:

- the three regional database management systems (DBMS),
- a web-mapping and geodata viewing application,
- the Study website,
- a Study-wide document management system with web interface, and
- a web-enabled version of the Shared Vision Model (SVM).

Given the distributed nature of the Study Participants and the stakeholders across the study region, the Internet should serve as the backbone for integrating the Study's IM system. Study web pages and hyperlinks contained therein then serve as the means for providing linkages among the recommended applications as well as the collection of documents, databases, images, etc. that comprise the Study's body of data and information. Hyperlinks to a particular document, database, or application should be present at all logical locations within the system.

Under this scenario, the Study website serves as the focal point, and point of departure, for all system functions. Organization of data and information services via the existing Study website will allow for the efficient and simple query and transfer of information to both the public and to Study Participants. Moreover, by utilizing the familiar structure of web portals as a central information store, all users of the system will immediately be able to find the information that they are seeking.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	i
TABLE OF CONTENTS	x
TABLE OF FIGURES.....	xiii
1.0 INTRODUCTION.....	1
2.0 EXISTING STUDY POLICY	4
2.1 Study Mandate	4
2.2 Plan of Study (POS)	4
2.3 Great Lakes Water Quality Agreement (GLWQA)	5
3.0 NEEDS ASSESSMENT	6
3.1 Overview of Needs Assessment Process	6
3.2 NAQ Responses and Follow-up Interviews	6
<u>3.2.1 Common Data Needs TWG</u>	<u>6</u>
<u>3.2.2 Hydrologic and Hydraulics TWG</u>	<u>6</u>
<u>3.2.3 Hydropower TWG.....</u>	<u>7</u>
<u>3.2.4 Coastal TWG.....</u>	<u>7</u>
<u>3.2.5 Environmental/Wetlands TWG.....</u>	<u>8</u>
<u>3.2.6 Recreational Boating and Tourism TWG.....</u>	<u>8</u>
<u>3.2.7 Study Board – Scirmemmano</u>	<u>9</u>
<u>3.2.8 Public Interest Advisory Group.....</u>	<u>9</u>
<u>3.2.9 Plan Formulation and Evaluation Group.....</u>	<u>10</u>
3.3 Study Properties and Needs	11
4.0 EXISTING RESOURCES AND POTENTIAL SYSTEM COMPONENTS /	
PARTICIPANTS.....	14
4.1 International.....	14
<u>4.1.1 Open GIS Consortium (OGC).....</u>	<u>14</u>
<u>4.1.2 Great Lakes Information Network (GLIN)</u>	<u>14</u>
<u>4.1.3 Binational.net</u>	<u>15</u>
4.2 United States.....	15
<u>4.2.1 Federal Geographic Data Committee (FGDC) & the National Spatial Data</u>	
<u>Infrastructure (NSDI)</u>	<u>15</u>
<u>4.2.2 Cornell University Geospatial Information Repository (CUGIR)</u>	<u>16</u>
<u>4.2.3 New York State GIS Clearinghouse.....</u>	<u>16</u>
4.3 Canada	16
<u>4.3.1 GeoConnections</u>	<u>16</u>
<u>4.3.2 Ontario.....</u>	<u>17</u>
<u>4.3.2.1 Land Information Ontario.....</u>	<u>17</u>
<u>4.3.3 Quebec.....</u>	<u>18</u>
<u>4.3.3.1 Environment Canada Quebec Region</u>	<u>18</u>

5.0 OTHER ORGANIZATIONS / STUDIES	19
5.1 Red River Basin Decision Information Network.....	19
5.2 Yellowstone-to-Yukon	20
5.3 United States Global Change Research Program's Data Working Groups.....	21
6.0 FOCUS DISCUSSIONS AT IMS WORKSHOPS.....	22
6.1 Policy	22
6.2 Technical.....	23
6.3 FGDC (USGS) Cooperative Agreements Program (CAP) Grant Opportunity.....	24
6.4 Study Properties and Needs	26
7.0 PRIMARY ALTERNATIVES AND OPTIONS.....	28
7.1 Data Discovery.....	28
<u>7.1.1 Alternatives</u>	<u>30</u>
7.1.1.1 Status Quo	30
7.1.1.2 Generating a Data List.....	31
7.1.1.3 Metadata Catalog.....	31
7.1.1.4 Participation in SDI	33
<u>7.1.2 Additional Options</u>	<u>34</u>
<u>7.1.3 Evaluation of Alternatives.....</u>	<u>35</u>
<u>7.1.4 Costs</u>	<u>37</u>
<u>7.1.5 Recommendations</u>	<u>39</u>
7.2 Data Storage, Access, and Distribution.....	40
<u>7.2.1 Alternatives</u>	<u>42</u>
7.2.1.1 Status Quo	42
7.2.1.2 Single Repository	43
7.2.1.3 Single Data Base Management System (DBMS).....	44
7.2.1.4 IJC Distributed DBMS	44
7.2.1.5 Regionally Distributed DBMS	45
7.2.1.6 TWG Distributed DBMS.....	47
<u>7.2.2 Additional Options</u>	<u>49</u>
<u>7.2.3 Evaluation of Alternatives.....</u>	<u>49</u>
<u>7.2.4 Cost.....</u>	<u>58</u>
Status Quo.....	58
Proprietary	60
<u>7.2.5 Recommendations</u>	<u>60</u>
7.3 Document and General Information Management Tools	62
<u>7.3.1 Simple Web Site Approach</u>	<u>62</u>
7.3.1.1 Hierarchical Web Page Structure	62
7.3.1.2 Optional Web Tools	64
7.3.1.2.1 <i>HTML Text Search</i>	64
7.3.1.2.2 <i>Metadata for aspatial information</i>	67

7.3.2 Document Management Systems	68
7.3.2.1 COTS Software	68
7.3.2.1.1 DocuShare – Xerox Corporation.....	68
7.3.2.1.2 EasyDocs - Internet Development Ltd., UK	71
7.3.2.2 Customized Information Management Systems.....	74
7.3.2.2.1 US EPA Environmental Information Management System	74
7.3.3 Presentation Options for Study Content and Decision Support	77
7.3.3.1 Stand Alone	77
7.3.3.2 Web Enabled	80
7.3.4 Recommendations	80
7.4 System Component Integration	81
8.0 IMPLEMENTATION	84
8.1 Data Discovery.....	84
8.2 Data Storage, Maintenance, Access, and Distribution.....	86
8.3 Document and General Information Management Tools	86
9.0 SUMMARY	87
10.0 REFERENCES.....	101
Appendix I: Needs Assessment Questionnaire distributed to all TWGs	102
Appendix II: Lists of model inputs and outputs	107
Appendix III: DIWG Policy Examples	117
Appendix IV: FY2002 CAP Grant Proposal Summary	118
Appendix V: Public Participation Management Tools	119
A.V.1 Information Collection Using Web-based Forms	119
<u>A.V.1.1 Web-Based Surveys.....</u>	<u>119</u>
A.V.1.1.1 HTML Survey Design	120
A.V.1.1.2 COTS Software Programs	122
<u>A.V.1.2 Feedback Forms.....</u>	<u>125</u>
A.V.2 Contact Addresses/Lists.....	128
Appendix VI: List of Acronyms	129

TABLE OF FIGURES

Figure 7.1.1 – Geospatial Data Discovery through the SDI	29
Figure 7.1.2 - Evaluation of Data Discovery Alternatives.....	37
Figure 7.2.1(a) - Flow of information between modeling groups and Single DBMS server in Alternative 3.....	50
Figure 7.2.1(b) - Flow of information between modeling groups and IJC Distributed DBMS servers in Alternative 4.	51
Figure 7.2.1(c) - Flow of information between modeling groups and Regionally Distributed DBMS servers in Alternative 5.	52
Figure 7.2.1(d) - Flow of information between modeling groups and TWG Distributed DBMS servers in Alternative 6.	53
Figure 7.2.2 - Evaluation of Storage, Maintenance, and Access Alternatives.....	57
Figure 7.3.1 - LMPDS Document Clearinghouse Contents Page.....	63
Figure 7.3.2 - Basic Search Form Example	65
Figure 7.3.3 - Query Results Page	66
Figure 7.3.4 - DocuShare Integration with Windows Explorer	69
Figure 7.3.5 - DocuShare Outlook Integration	69
Figure 7.3.6 - EasyDocs Search Screen Example.....	72
Figure 7.3.7 - EasyDocs Search Results Example.....	72
Figure 7.3.8 - EIMS Search Form.....	75
Figure 7.3.9 - EIMS Metadata Search Results List	75
Figure 7.3.10 - EIMS Metadata Summary Form	76
Figure 7.3.11 - The Microsoft Encarta Start-Up Window	78
Figure 7.3.12 - Zambezi River Information management System GUI	79
Figure 7.4.1 - System Components Integration	83
Figure 8.1.1 - Alternatives and Options Timeline.	85
Figure A.V.1 - Example of Web Based Questionnaire.....	121
Figure A.V.2 - HRDC NOC/SIC Code Database Query Screen	122
Figure A.V.3 - Feedback / Submittal Form for Section 227 Project Information	126
Figure A.V.4 - Section 227 Database Query Form.....	127
Figure A.V.5 - Section 227 Search Results Page.....	127

1.0 INTRODUCTION

The Common Data Needs Technical Working Group (CDNTWG) of the International Joint Commission's Lake Ontario – St. Lawrence River Study (LOSLR "Study") was charged with the development and implementation of an Information Management Strategy (IMS) for the Study. In response, the CDNTWG assembled an IMS Team consisting of GIS, IM, and IT professionals either participating in the Study or associated with agencies or organizations in the Study region. With assistance from a contractor, Pangaea Information Technologies, the IMS team has conducted a comprehensive Needs Assessment and conducted two workshops to aid in the formulation of that Strategy. The results of that development effort are presented in this document. After selection of implementation alternatives and options, and endorsement of Study IM policies to support those alternatives and options, the CDNTWG will coordinate the development and implementation of a detailed Information Management Plan.

Information management is a critical component of any study conducting a regional impact assessment with as large and diverse a scope as the Lake Ontario – St. Lawrence River Study. At the heart of the study's results will be the data and information collected, analyzed and produced by the Technical Working Groups (TWGs) upon which decisions will be made and justified. To ensure the impartiality of the study's conclusions, ideas and information will need to be exchanged freely and openly among study participants and in as near real-time as possible. For the LOSLR Study, an information management system will be required to organize a large amount of geospatial and non-spatial data. The procedures and mechanisms employed in such a system will need to facilitate the sharing of ideas and information to a distributed set of users. Because the results of the Study will formulate recommendations that could affect large segments of the population, an information management system developed for the Study should address the public's need for free and open access to information.

One of the principle functions required of a system responsible for supporting the sharing of information is the ability of potential users to learn of the existence of and significant details about data or information. Popularly known as *data discovery*, such mechanisms employ search procedures which match user defined criteria against information held about the data, known as metadata. *Metadata* is the most important component of the data discovery process. The degree to which an organization generates metadata for data and information determines how effective a data discovery mechanism can be implemented. Standard geospatial metadata formats have been developed by the FGDC and ISO to ensure that all essential information about the data has been collected and represented in a consistently organized and searchable way. Metadata that is not compliant with commonly accepted standards lacks the necessary completeness that is

required to publish information in a meaningful way. Because data discovery only provides information contained in metadata, it can be considered separate from most liability and security concerns associated with data access and distribution.

Data storage, maintenance and access needs require the coordination and integration of the many responsibilities associated with the system and its data. *Data owners* hold the responsibility for data security, use and maintenance, and they have the authority to define and manage data access and distribution. This is commonly achieved through the application of a flexible security model. *Data stewards* are responsible for the day-to-day maintenance of data. Given the authority by the data owner, data stewards are familiar with the issues and concerns specific to a data set. Consistent maintenance is essential to ensure the currency and quality of data. The integration of these responsibilities with the infrastructure and organizational procedures that support the system ensures reliability and sustainability.

The Internet continues to be the most commonly utilized method of distributing an information management system to a large number of dispersed users. Through consistent and well-designed implementation, an information management system can consist of one or many servers and provide simultaneous access to multiple users in many different locations. In establishing a shared environment from which data and information resources can be utilized, a client/server strategy can promote efficient system administration and access. The organization and design of the system will also need to address extensibility in terms of supporting the capacity to implement additional technical functionality after the initial implementation phase. Examples of such additional functionality are web services such as *web mapping services* (WMS) and *web feature services* (WFS) which provide functionality for interactive geospatial data viewing and querying over the Internet. This type of functionality would require the implementation of a system that allows for connectivity to multiple datasets, potentially over multiple systems.

Developing a system that includes the required functionality begins with the strategy and analysis phase of the *system development life cycle*, the traditional methodology used to develop information systems. The primary purpose of the strategy and analysis phase is to establish a solid understanding of the organization and functions for which the information system is being developed. Done through research, conducting needs assessment, and interviewing relevant participants and system users, this process leads to a strategy appropriate to a particular organization with its specific needs. The strategy will provide guidance through following phases of the system development life cycle. This report represents a synthesis of the information collected during the strategy and analysis phase of system development for the IJC LOSLR Study.

Through collecting more detailed information about the specific data for which the system is being developed, logical and physical models will be developed to provide a clear concept of the system and direct the build and document phase of the development process. The initial implementation of the system will require the installation of hardware and software, loading of sample data, and the development of user documentation, help text and operations manuals to support the use and operation of the system. With initial implementation complete, the testing phase of the development process will ensure the system performs the required functions and supports the organizations business processes as they were defined in the strategy and design phases. After testing has been completed and the system refined, the production phase, the final step of the development process, can be initiated. The final system is delivered to the users and any necessary training is conducted. The system is closely monitored through the early period of production and enhanced or refined accordingly to optimize system performance.

This report is intended to summarize the strategy and analysis phase of the system development process and provide appropriate alternatives from which a specific system design and implementation plan can be formulated. This report is not intended to provide the design specifications required for building the system. However, through the alternative approaches discussed in this report, direction is provided to assist in the beginning of the design phase of the development process.

2.0 EXISTING STUDY POLICY

Policies with organizational authority are necessary to ensure consistency throughout the Study and support the coordinated effort required for success. Formal policies, in the form of directives or mandates, provide clear guidance and can serve as a model for other Studies. The needs assessment revealed a very limited number of IJC-level policies on data management, and a few more at the Study-level, though relatively general in most cases.

2.1 Study Mandate

The following policy statements are taken directly from the Mandate for the IJC LOSLR Study.

10. “The Commission emphasizes the importance of public outreach, consultation, and participation. ... The Commission expects the Study Board to involve the public in its work to the fullest extent possible. The Study Board shall provide the text of media releases to the Secretaries of the Commission prior to their release.”

11. “To facilitate public outreach and consultation, the Study Board shall make information related to the study as widely available as practicable, including white papers, data, reports of the Study Board or any of its subgroups, and other materials, as appropriate.”

2.2 Plan of Study (POS)

The following policy statements are taken directly from the Plan of Study for Criteria Review for the IJC LOSLR Study.

4. Coordination of Common Elements by the Study Board

4.1 Direct and Coordinate Work of Study Teams

4.1.d The authority and tasks of the Board would include to “act as coordinator to ensure effective exchange of information among the study teams, and full use of studies or information from other sources.”

4.6 Process Management and Integration of Work

“Given the considerable cost of the overall Plan of Study activities, the Study Board will also need to ensure that duplication of effort is minimized, and data collected is made widely available across all teams.”

“The Study Board will also need to satisfy itself that each Study Team is carrying out the required work in a satisfactory manner, and that cross-interest impacts have also been considered.”

Annex 4 – Background Documentation and Correspondence

4(c) Directive in the Lake Ontario – St. Lawrence River “Plan of Studies” Team

“Documents, letters, memoranda, and communications of every kind in the official records of the Commission are privileged and become available for public information only after release by the Commission. The Commission considers all documents in any official files that the team may establish to be similarly privileged. Accordingly, all such documents shall be so identified and maintained as separate files.”

2.3 Great Lakes Water Quality Agreement (GLWQA)

The policy statements listed below are taken from the 10th Biennial Report on the Great Lakes Water Quality, Chapter 6 Information and Data Management, and are only directly applicable to the GLWQ Agreement.

- Quality assurance for legal and scientific defensibility
- Broad waiver of data recovery costs
- Promote accessibility of data and information
- Promote organization and management of data bases
- Establish protocols to ensure compatibility and comparability of data [across programs and boundaries]

3.0 NEEDS ASSESSMENT

3.1 Overview of Needs Assessment Process

The Needs Assessment (NA) process consisted of the development of a Needs Assessment Questionnaire (NAQ; presented as Appendix I), completion of the NAQ by TWGs and the Public Interest Advisory Group (PIAG), and follow-up interviews conducted by Pangaea. A list of datasets associated with the inputs and outputs of models addressing Performance Indicators (PIs) was compiled as a result of this and associated efforts, and is presented as Appendix II to this report.

3.2 NAQ Responses and Follow-up Interviews

Responses to the NAQ were received from all of the TWGs except two: Commercial Navigation and Municipal, Industrial, and Domestic Water Use. Follow-up interviews were conducted via conference call after reviewing the completed questionnaires.

3.2.1 Common Data Needs TWG

The Common Data Needs (CDN) TWG was formed to provide for the elevational (bathymetric and topographic) data and imagery requirements for the compliment of TWGs, and to work towards an information management strategy to facilitate the sharing, access and use of all data and information generated within the study. The CDN TWG is nearing completion of data collection activities for the elevation and imagery data, as well as geodata for shorelines, political units, transportation features, watersheds, and tributaries. All the data will need to be transmitted to the other TWGs for use in their modeling and analysis. An FTP site is expected to be sufficient for these data transfers through the summer 2002, though this does not allow for data viewing. The group has identified some restrictions on data use (e.g., with the City of Kingston and City of Hamilton orthoimagery).

3.2.2 Hydrologic and Hydraulics TWG

The Hydrologic and Hydraulics (H&H) TWG will produce a series of hydrologic scenarios that describe the levels and flows associated with different regulation plans and climate conditions. These scenarios will be used to perform resource-level impact assessments: the TWGs (other than CDN) will use these in models which address the PIs associated with their particular resource sector. Hence, upon completion of hydrologic data set production, H&H TWG will need to make their products available to the TWGs.

The H&H TWG's response to the NAQ focused on data needed to support their modeling efforts, data transfer to other TWGs, and archiving their outputs.

H&H TWG model inputs and outputs will need to be made accessible to the public as well as the other TWGs. However, with respect to the outputs, the H&H perceives that the usefulness and desirability of the *full* datasets to the public is minimal. The only input datasets not accessible to the public are a portion of the raw digital bathymetry and raster shoreline files for the Upper St. Lawrence River. These are owned by Canadian Hydrographic Service (via Nautical Data International) and were made available to Environment Canada for modeling purposes only.

Upon completion of dataset production, at which point updates will no longer be needed, the H&H TWG believes it best to transfer all data to a single repository. Prior to summarizing, there will be about 400 Mb of output for each scenario for each geographic location within the Study area that is being modeled.

3.2.3 Hydropower TWG

The Hydropower TWG is still defining Performance Indicators (PIs), models, and thus input data needs. Some data needs for the group have been addressed: as "Hydropower Entities" associated with the TWG hold a wealth of historical data that is already in the public domain. These data sets are maintained by the Entities themselves, and are continually being updated. At the time of the interview, the Hydropower TWG believed that all of their data could be shared with other groups, with the exception of some megawatt pricing information. However, the Entities would not want to expend the resources to actively maintain and update their datasets in a remote location on a frequent basis. This group perceives a need for consistent, Study-wide guidelines for PI valuation.

3.2.4 Coastal TWG

The Coastal TWG is split into two sub-groups: one for Lake Ontario and the upper St. Lawrence River, and the other for the Lower St. Lawrence River. Both groups have defined their PIs, modeling approaches, and input data needs. Their model inputs and geospatial data requirements are extensive relative to the other TWGs (see Appendix II). The "Upper" sub-group of the Coastal TWG will use Baird's Flood and Erosion Prediction System (FEPS), with data processing done primarily by consultants. The TWG is purchasing a Coastal Data Server (CDS) for the Upper sub-group, which will hold all of their data. The sub-group considered the possibilities of holding all data for the Study on this server, though they do not currently have the funding necessary for public accessibility to the CDS. The "Lower" sub-group, centered at the Environment

Canada, Meteorological Service of Canada, Quebec Region, Hydrology Section (ECQR), is utilizing an approach relying on finite element grid model output. A new database management system (DBMS) is being developed at ECQR, in part to address the Lower sub-group's needs.

The Coastal TWG has some liability concerns associated with premature release of data that might allow for misuse and misinterpretation of the data. Also, there are potential security issues with some higher-resolution data sets (e.g., aerial photos, topometry). The Coastal TWG noted that some datasets are licensed and owned by private companies. The group perceived no problems with the short-term GIS Guidelines.

3.2.5 Environmental/Wetlands TWG

The Environmental/Wetlands TWG provided a limited response to the Needs Assessment Questionnaire. The group has identified an initial set of Performance Indicators, models, and supporting data inputs. The data modeling and processing will be performed by a combination of affiliated organizations, consultants, and TWG members. Some of their data will be funded and owned by the Ministry of Natural Resources. The group may require additional GIS capabilities and technical support. A need for designation of an information management lead for each TWG was suggested. The sole responder to the NAQ suggested that it was likely that an FTP site could meet the group's needs for information distribution. At the March 7th and 8th Environmental/Wetlands TWG meeting, it was confirmed that providing data discovery, evaluation (including visualization), and access within this TWG and among others could be of considerable benefit.

3.2.6 Recreational Boating and Tourism TWG

The Recreational Boating and Tourism TWG has identified their Performance Indicators, models, and data inputs. To date, much of the information identified as "required" by this TWG is survey-based (data about marina owners and recreational users). As a result of this, the group has confidentiality concerns (e.g., individual survey responses). This TWG also needs highly resolved bathymetric data for areas around the marinas, boat docks, and launching ramps. The geodata inputs to the models were created in UTM, and will be reprojected as per the *Short-Term GIS Guidelines*. Aside from this, the group has not yet evaluated the *Short-Term GIS Guidelines*, including the metadata standards. The group does not have funding for public awareness and access included in their budget, and believes that the policy and funding for public accessibility and data sharing should come from the IJC -- because they are the principal owners of the data during the course of the Study. The group does see that there could be potential benefits from learning

about the other TWGs' model inputs and outputs. They foresee that extending applications developed for the Study could benefit other user groups during or after the Study.

3.2.7 Study Board – Scirmemmano

Frank Scirmemmano had several recommendations for the Study, most relating to information accessibility and transparency with respect to the public. Scirmemmano believes that public awareness and accessibility should be conducted at all organizational levels with consistent methods and structures. In general, all the data should be made available for public scrutiny. This is necessary to ensure transparency of the Study process, which is crucial to public acceptance and perceived credibility of the Study. Hence, access is important for success of the Study. In order to address public accessibility, there is a need for a Study policy regarding what data and information should be made accessible to the public. Scirmemmano generally approves of the draft definition of “publicly-accessible data”: new data, or value-added data produced by the Study that is not readily-available through other sources.

Scirmemmano also related the need for Study-wide communication in the area of data collection and use. The discovery, acquisition, and use of data should be coordinated to maximize the efficient use of Study resources. Also, the development of the IMS strategy is crucial to the success of the Study. In order to ensure compliance, contracts and the balance of funding should be tied to compliance with the process and related policies that make up the IMS strategy. However, some contingency should be created so that compliance to communication, metadata, and data policies, despite their importance, do not detract from the funding obligations to the Study's working groups (or inhibit the TWGs from completing their research or analyses).

3.2.8 Public Interest Advisory Group

The Public Interest Advisory Group's (PIAG) response to the Needs Assessment focused on the need for public accessibility and transparency. The group's principal concern is in maximizing the public's knowledge of the Study, as well as facilitating the public's involvement in the Study process. As knowledge and participation increase, the public's perception of the Study's credibility increases. This perception of credibility is crucial to the success of the Study.

The first step in facilitating public involvement is to advertise the Study's existence to the public. PIAG plans to increase visibility of the Study through press releases, status updates and reports. In order to address the need for public accessibility and

transparency, PIAG recommends that all data, models, procedures, and policies be disclosed throughout the duration of the Study, and for some time thereafter. Also, in order for the public to access the data, it needs to be available to them at the appropriate level of complexity. Therefore, the level of detail and subject matter for all data, including summary reports and periodic updates of TWG activities should correspond to the needs and desires of the target audience. This disclosure has the potential to be an involved process requiring a system capable of managing public surveys and contact address lists.

Along with assuring that information is matched with the audience at the appropriate level of detail, the information must be organized in a manner that enables public consumption. To best aid the public, the information will be organized with an emphasis on usability in terms of content and format, via a hierarchical information structure. A document search and retrieval functionality must exist, which can be implemented in several different ways, ranging from documents searched and presented through metadata search functionality (such as the RRBDIN web site) to a hierarchical web page structure where the public can pass through hyperlinks for increasing detail. The public also needs the ability to find specific information within Study databases. To facilitate this, there needs to be database search and display functionality. There are many options for this search and display functionality, ranging from a stand-alone database with query and report functions to a web-enabled query and reports system associated with a backend database. A minimum requirement will be the capability to navigate through the information using some form of HTML text search capability.

To ensure that the system will meet public needs and consider public input, the information access structure also needs to allow for public feedback and/or questions that can be considered and responded to in a timely manner. In certain cases, it may be necessary to direct such questions to one or more appropriate Study members. However, a separate ask-an-expert capability, in which questions would be forwarded to experts within the various TWGs, was discouraged by the PIAG during their interview.

3.2.9 Plan Formulation and Evaluation Group

The Plan Formulation and Evaluation Group (PFEG) consists of all Study leaders: the Planning Group members, the entire the Study Board, and a representative (Co-Chair) from each of the technical work groups (TWG), and the Public Interest Advisory Group (PIAG). PFEG needs to be able to receive the output from the TWG models, each addressing a Performance Indicator response to the different hydrologic regimes (levels and flows) specified by the H&H TWG. At present, this information – presented in terms of monetary gains/losses inasmuch as possible – is to be integrated through Shared

Vision Planning. A specific Shared Vision Model will be developed for the Study. This decision support tool will assist stakeholder groups in the comparison of alternative regulations and their associated hydrologic regimes and resource sector impacts.

3.3 Study Properties and Needs

The Needs Assessment process revealed information essential to the formulation of an information management strategy. First, two primary user groups associated with the Study were recognized: Study Participants (all TWGs, the PIAG, and the Study Board), and the Public. Second, the general flows of information in the Study were identified. Third, there were requests, suggestions, comments, and needs identified that were either held in common among the individual Study Participant groups, or should be considered at the Study-level.

The general flows of information in the Study start with regulatory alternatives passed to the Hydrologic and Hydraulics (H&H) TWG. The H&H TWG models the “levels and flows” scenario associated with each regulatory alternative, and provides these to the TWGs who evaluate resource sector response for selected Performance Indicators (PI). In addition to these “levels and flows”, which can be considered forcing or driving variables of the PI models, modeling groups within the TWGs require many other input variables. While many model inputs are derived directly from organizations outside of the Study, inputs from at least three sources inside the Study can be used:

- 1) “basemap layers” provided through the CDN TWG,
- 2) model inputs obtained from other TWGs, or from modeling teams within the same TWG, and
- 3) model outputs from other TWGs, or from modeling teams within the same TWG.

All model outputs, aggregated as appropriate and valued in dollar (\$) terms whenever possible, will be made available to all Study Participants and the Public. As suggested above, some of these model outputs may serve as inputs in models addressing PIs in the same or different resource sector. Model approach, results, and analysis, and discussion will be documented in report form, which will be made available to all Study Participants, and to the Public via the PIAG. Last, all model outputs will be transferred to the PFEG for incorporation in the Shared Vision Model, subsequently used by all Stakeholders.

These flows of information can be summarized as follows:

Primary Model Drivers

H&H TWG → TWGs other than CDN

Other Model Inputs

Non-Study Orgs. → TWGs

CDN TWG → other TWGs

TWGs → TWGs other than CDN

Model (Outputs) Results

TWGs → Model Inputs for TWGs (other than CDN)

TWGs other than CDN → PIAG → all Study Participants and the Public

TWGs other than CDN → PFEG → Stakeholders (via Shared Vision Model)

These study-wide results can be summarized as “Study Properties”, and as “Study Needs”:

Study Properties

- Modeling and data processing is widely distributed within the TWGs among contractors, affiliated agencies and contractors.
- TWGs identified specific datasets as sensitive (for reasons related to security, proprietary, and liability).
- Inter-TWG data discovery (and access & evaluation) mechanism, while not currently in place, has potential benefits for input/output evaluation, further development of PI models, and better overall integration in the Study.
- The need for inter-TWG data transfer is currently limited. Provision of a mechanism for comprehensive data discovery, evaluation, and access would increase the need for inter-TWG and intra-TWG data transfers.
- The Common Data Needs TWG’s *Short-term GIS Guidelines* were consistently accepted by those TWGs that responded to the NA questionnaire.
- Most TWGs stated that obtaining metadata would be relatively easy, but none had it in a format ready for the Study.
- Responses to questions about making activities, forms or procedures web-enabled involved concerns of funding.
- Funding for public accessibility of Study information is not presently budgeted at the TWG-level.

Study Needs

- Information management strategy that will provide for a widely distributed user group.
- Flexible security model and communication of data set sensitivity to users (via metadata).
- Data discovery, evaluation, and access mechanism.
- Policy and mechanism for data archiving.
- More specific metadata guidelines for TWGs.
- Importance of making the process transparent to the public.
- Study policy on public accessibility to data and information.
- Study policy on bilinguality of metadata and/or data.
- Study policy or clarification on PIAG vs. TWG responsibility and funding for public access and outreach.
- Study policy regarding what data and information should be made accessible to the public (e.g., “new and value-added not otherwise readily-available through other sources”).
- Study policy on tracking sensitive data (including that which is licensed).
- Study policy that links compliance to metadata and data standards to contracts and the balance of funding.

[Please note that many properties and needs directly associated with PIAG activities are not included above, though some options will be addressed in Section 7.3 and Appendix V.]

Fourth, a substantial number of both geo-spatial and non-spatial databases were identified as inputs and outputs as related to Study Participant activities. Most of these were associated with TWG models that address PIs. A list was generated from the Needs Assessment. For those TWGs with limited (or no) response to the NA questionnaire, the Plan of Study was used to augment (or wholly create) their parts of the input/output list. This list, comprehensive as possible at the time of this report, is presented in Appendix II.

4.0 EXISTING RESOURCES AND POTENTIAL SYSTEM COMPONENTS / PARTICIPANTS

The Common Data Needs TWG held an Information Management Strategy Workshop in Burlington, Ontario February 14th and 15th. The workshop consisted of three presentations, followed by focused discussions. The first presentation focused on the results of the NA process. The second focused on existing IM resources (systems, knowledge bases, etc.) available to the Study, as well as IM strategies, policies, and “lessons learned” by organizations as structurally- and functionally-similar to the Study as possible. The third presentation explored potential policies, and alternative system architectures to meet Study needs. “Break-out groups” engaged in focused discussions on IM policy, technical issues, and writing a proposal for funding assistance to help meet Study IM needs. The latter topic focused on a FY2002 Category 4 CAP Grant, funded jointly through GeoConnections (Canada), and its US counterpart, the Federal Geographic Data Committee (U.S Geological Survey). [A grant proposal was submitted and has been accepted. See Appendix IV for a summary of the proposed project.]

4.1 International

4.1.1 Open GIS Consortium (OGC)

The Open GIS Consortium (OGC) is a consortium of government agencies, non-profit organizations, universities, and private organizations working together to ensure interoperability in the geospatial community. The group is working towards interoperability by developing standards for data formats and quality and procedural standards. The Consortia is working to develop and incorporate ISO standards so that all geodata and services can be internationally compatible. Through the widespread use of these standards, the Web can become “geo-enabled,” which will allow geospatial data to be more widely used and therefore incorporated into more decision-making processes.

4.1.2 Great Lakes Information Network (GLIN)

The Great Lakes Information Network (GLIN) serves as the clearinghouse of Great Lakes information. It was established by the Great Lakes Commission and has been online since 1993. The Great Lakes Commission was established in 1955 by federal legislation as an interstate agency to work with the eight Great Lakes States in the United States. GLIN now serves as the gateway for Great Lakes geospatial data, and provides some Web Map Services. GLIN is structured as a decentralized network of regional partners and information providers (including USEPA, USACE, IJC, and Environment

Canada). GLIN provides centralized access to the information providers via a page of links to the individual organizations. The information providers develop, host, and maintain their information at their own location. GLIN serves to organize and enhance access to the information by providing a central link to all information in the network. This works to increase exposure of the information through integration among the information providers in the context of the partnership network. GLIN will soon serve as the site for the GLIN Data Access Clearinghouse (GLINDA). GLINDA will serve as a clearinghouse for all GLIN data and will be a node of the Federal Geographic Data Committee (FGDC) National Spatial Data Infrastructure (NSDI), facilitating more widespread data discovery.

4.1.3 Binational.net

There are many binational program web sites that currently are hosted on Canadian and United States web sites. This split between country web sites makes the discovery and access of data difficult for the users of the binational data. Until this program, it has been difficult to create a system capable of hosting both countries' data due to different regulations in each country governing website design. Binational.net could get around these regulations by starting a website that is not under the auspices of either country; having a .net address rather than .ca or .gov. Binational.net was announced at an IJC conference and is a joint venture by the EPA and Environment Canada - Ontario Region that seeks to eliminate the redundancy and confusion created with multiple hosting sites by hosting binational data for the United States and Canada at a single location. The focus for this project is in making governmental data easily accessible to the public.

This option is of potential interest to the Study, but it appears that some procedural steps, at least from the US side, have slowed the development of binational.net. One criticism thus far is difficulty using the system due to inefficiencies in speed. More information is currently needed in order to fully consider the potential benefits of this resource.

4.2 United States

4.2.1 Federal Geographic Data Committee (FGDC) & the National Spatial Data Infrastructure (NSDI)

The USGS's Federal Geographic Data Committee (FGDC) is made up of 17 federal agencies to promote the nationwide use and sharing of geospatial data. The FGDC, with the help of other partner organizations from state and private entities, is the developer for the National Spatial Data Infrastructure (NSDI). The NSDI sets standards for data and

metadata in terms of both quality and format. The FGDC also maintains a network of decentralized metadata clearinghouses which users can query to find needed metadata based on keywords, time period, and/or geographic location. The FGDC is also involved in data, metadata, and infrastructure development by offering funding opportunities via Cooperative Agreements Program (CAP) grants.

4.2.2 Cornell University Geospatial Information Repository (CUGIR)

Cornell University Geospatial Information Repository (CUGIR) is an active online repository providing geospatial data and metadata for New York State, with special emphasis on those natural features relevant to agriculture, ecology, natural resources, and human-environment interactions. Subjects such as landforms and topography, soils, hydrology, environmental hazards, agricultural activities, wildlife and natural resource management are appropriate for inclusion in CUGIR. All data files are cataloged in accordance with FGDC standards and made available in widely used geospatial data formats.

4.2.3 New York State GIS Clearinghouse

The New York State GIS Clearinghouse has a number of primary functions. It serves as an access for (downloading) data associated with the state of New York at no cost to users for some limited datasets. It also houses the New York State GIS Data Sharing Cooperative. The New York State GIS Data Sharing Cooperative is a group of government agencies and non-profit organizations who have entered into Data Sharing Agreements. When groups have entered the Cooperative, they must provide metadata to the clearinghouse, and also fill GIS data sharing requests from other members of the cooperative. Members are encouraged to put their data on a web site to minimize the need for staff interaction. There is no cost for joining, aside from costs associated with meeting member requests for data.

4.3 Canada

4.3.1 GeoConnections

GeoConnections is a national public and private partnership initiative led by Natural Resources Canada (NRCAN), and serves a role generally analogous to the US FGDC. The initiative fosters the creation of a Canadian Geospatial Data Infrastructure (CGDI) to enable online access and sharing of geographic information and services. GeoConnections is jointly funded by the Canadian government and partner organizations.

GeoConnections provides many services to fulfill the needs of its different users. The GeoConnections Discovery Portal (formerly CEONet) allows users to search available metadata to discover data. Metadata is reviewed initially upon receipt from data owners/distributors, and periodically thereafter, by the Metadata Content Team. GeoConnections offers support for Web Mapping Services (WMS) and Web Feature Services (WFS), which utilize Open GIS Consortium (OGC) interfaces. GeoConnections also offers CGDI Re-Usable Components (e.g., Earthscape map viewer clients), which are a set of tools that provide geospatial location display in Web pages. Standardized interfaces (wizards) are provided for Re-usable Components so developer can embed these tools within their own web-based application. GeoConnections also permits the use of Web API, which enables customized portals into any part of the GeoConnections Discovery Portal, giving any external website the capability to use any CEONet service. GeoConnections also plays a role in data development by offering funding opportunities via “Access” grants, CAP grants (a cooperative effort with the FGDC), etc.

4.3.2 Ontario

4.3.2.1 Land Information Ontario

Land Information Ontario (LIO) was designed to create a common infrastructure that will allow a wide range of consistent and well-managed land information to be captured, cataloged, and made readily available from a centralized warehouse. The group has coordinated Ontario’s participation in the CGDI.

There are several components within LIO. The Ontario Land Information Warehouse (OLIW) allows for online data viewing using the OLIW Map Browser. The Warehouse contains 140 viewable geospatial datasets, however, the data cannot be extracted and downloaded by users, except by subscription. The Ontario Land Information Directory (OLID) allows for data discovery by keyword and geographic area. The Ontario Digital Geographic Database (ODGD) consists of datasets maintained by the Ontario Ministry of Natural Resources (OMNR) that contains OMNR base data plus features of interest to the OMNR. Public and private users can access the data via several different licensing options. Users can purchase an Electronic Intellectual Property Copyright License if they have no plans to redistribute the data in any way. A Non-Value Added Resale License is available for users to sell and distribute the original data. Also, a Value Added Resale License is available for users to enhance, resell, and distribute the data. Finally, the Ontario Geospatial Data Exchange (OGDE) is a collection of shared data that only members can access. Membership is open to several different categories of organizations. All Schedule I and III ministries within Ontario are expected to join and must pay an annual levy of up to 50,000 CDN in order to join. Community groups with

an annual budget over 100 million pay no fee in the first year; and then 3,000 CND annually each following year. Community groups with an annual budget less than 100 million pay no fee in the first year; and then 1,000 CND annually each following year. Other nations and commercial groups are considered on a case by case basis. LIO is currently working with GeoConnections to incorporate OGC-compliant Web Feature Services and Web Mapping Services connectors to improve effectiveness of data use.

For each of the various datasets housed in the LIO warehouse, a custodian, or data steward(s), is identified by the data owner. The custodian has a support network of “information teams” who help define standards and protocols. The custodian is ultimately responsible for defining specific database updating and maintenance tools. LIO has carefully documented their policies, standards, and procedures and made this documentation available to the Common Data Needs TWG.

4.3.3 Quebec

As with Ontario, Quebec contains at least two existing resources for the LOSLR Study. These include the Quebec Ministry of the Environment (QME), and Environment Canada, Meteorological Service of Canada, Quebec Region, Hydrology Section (ECQR). The portion of the St. Lawrence River in Quebec is classified as an International Seaway. As such, the development and maintenance of a system for the hydrologic and coastal processes occurs at the federal level in Quebec. Hence, this is one reason that data development and modeling activities associated with this Study are being conducted with ECQR resources (i.e., facilities and staff). Although coordination with QME staff may be necessary and/or advisable at times during the Study, only ECQR resources will be described below.

4.3.3.1 Environment Canada Quebec Region

The Environment Canada, Meteorological Service of Canada, Quebec Region, Hydrology Section has recently made a substantial investment in a database management system: ~\$50,000 CND for hardware and software alone. This was purchased by EC for the purpose of EC activities with the understanding that much of the IJC LOSL Study’s coastal analysis of the lower St Lawrence would utilize the system. Currently in its implementation stage, the system is designed to run Oracle and support OGC-compliant geospatial web services. Experiences gained in the development, design and implementation of the information management system by EC staff and the Database Administrator constitutes a valuable resource for the Study and increases the available knowledge base (KB).

5.0 OTHER ORGANIZATIONS / STUDIES

Information management has been addressed by many organizations engaged in work of a comparable nature to the LOSLR Study. Many of these organizations were identified at the start of the IMS development process and have been evaluated in terms of their information management approaches. Lessons learned from the policies and decisions implemented by other organizations serve to promote a more thorough and successful information management strategy for the LOSLR Study. Organizations reviewed for their information management policies and decisions include the Yellowstone to Yukon Conservation Initiative, the Red River Basin Decision Information Network, Data and Information Working Group of the United States Global Change Research Program (USGCRP), and the Data Management Working Group of the USGCRP National Assessment Program.

5.1 Red River Basin Decision Information Network

The Red River Basin (RRB) Decision Information Network was developed to create an internet-based information dissemination system for the Red River Basin. The RRB Decision Information Network has two primary components, the RRB Decision Support System, and the RRB Virtual Data Base. Data planned for inclusion in the RRB Virtual Data Base are an authoritative base map for the basin, spatial data (e.g. topography, imagery), water quality data, and other related information. The International Joint Commission (IJC) is the data provider to the RRB Virtual Data Base. For the purposes of data discovery, the RRB Virtual Data Base is integrated with the Manitoba Land Initiative (MLI) in terms of shared web server and shared metadata catalog. The MLI ensures the long-term viability of the RRB Virtual Data Base by maintaining the metadata catalog. Data discovery queries to the RRB Data Information Network are processed on a replicate metadata catalog outside the Manitoba government firewall.

The RRB Decision Information Network recommends that a “watch-dog” group be created for the maintenance of the metadata catalog. This group would be responsible for overseeing the maintenance, collection, and integration of metadata from private organizations and governmental agencies in the RRB Virtual Data Base system on an ongoing basis. This “watch-dog” role would be carried out by maintaining contact with all private and non-government agencies that have contributed metadata in the past, and ensuring that metadata collection is current and that metadata is accurate. The group would also assist agencies with metadata collection tasks, help to identify new data sources related to Red River Basin flood management, and assist with integrating new and revised metadata.

The RRB Decision Information Network is also intended to facilitate future development of decision support tools via the RRB Decision Support System. The network is administered jointly by the IJC's Red River Task Force, and by the Global Disaster Information Network under the direction of the Office of the US Vice-President.

5.2 Yellowstone-to-Yukon

One component of the Yellowstone to Yukon (Y2Y) Conservation Initiative, the Y2Y Framework Dataset Demonstration Project, is a collaborative transboundary project focused on creating 10 seamless geospatial datasets from the best available sources. The Project is working with the FGDC and GeoConnections to develop the Framework Architecture needed to facilitate the transboundary sharing and use of the datasets. The Project has compiled a catalogue of FGDC-compliant metadata for project data. Quality Assurance/Quality Control has been a priority, checking to ensure attribute consistency, spatial accuracy, vertical integration with other project layers, and metadata completeness. The Project has also established policy to document errors and/or inconsistencies in the metadata. This facilitates a more informed data evaluation by potential users in relation to data quality and limitations. The Y2Y project also established clearinghouse nodes to allow for public data discovery. The group recognized the need for stable server location for the nodes. They were able to find these in the US, using agency and university resources. Nodes were also established in Canada, with the Canadian and US clearinghouses employing mirrored indexes. These nodes create a distributed, virtual warehouse from which the public can access data.

Along with node creation, the project has also established procedures for data storage, maintenance, and access location based on stewardship need. There is a three-tiered approach for data stewardship need. Core datasets (e.g. "those not likely to change) are mirrored on servers in Canada and the US. Metadata lists both site URLs in the "Distribution" section of the FGDC standard metadata (section 6). Datasets with routine updates (e.g. roads) remain under the stewardship of the owner. The owner provides maintenance and archiving offsite and the data is completely reloaded as needed. The third tier is for owners with access capacity who maintain their own data online or on-request.

For access, the project has a drill-down approach in HTML pages for larger, tiled datasets. They are developing specific protocols for updating databases. Y2Y uses a membership approach and agreement to an "acceptable use policy" to address access to data with significant licensing issues. The group has piloted the use of core data in a transboundary cumulative effects analysis application, and has plans to develop other applications.

5.3 United States Global Change Research Program's Data Working Groups

Two primary groups address data policy issues within the United States Global Change Research Program (USGCRP). The Data Management Working Group (DMWG) formulates data policy related to the diverse studies conducted under the auspices of the USGCRP's National Assessment Program. The Data and Information Working Group (DIWG) serves the same function for the USGCRP at large. The policies created by the DMWG include:

- 1) "Suggested Data Product Requirement for Grants, Cooperative Agreements, and Contracts" should be included in every contractual document (1997; see Appendix III),
- 2) Metadata should meet the FGDC standards (1998),
- 3) Servers be ANSI Z39.50 compliant (1998), and
- 4) Data abstracts should be submitted to the GCMD Global Change Master Directory (1998).

The Global Change Data and Information System (GCDIS), managed by the DIWG, provides a gateway to data and information related to global environmental change generated by federal agencies participating in the USGCRP. The GCDIS provides access to a wealth of documents related to data policies, including those advocated by the DIWG. The GCDIS is governed by the idea that there should be "full and open sharing" of data for free or at cost; via the World Wide Web whenever possible (DIWG 1991). To ease sharing, the GCDIS requires a standard data citation format (DIWG 1998; see Appendix III). The GCDIS also provides an important public outreach function through its "Ask Doctor Global Change" which puts users in touch with experts.

6.0 FOCUS DISCUSSIONS AT IMS WORKSHOPS

Three “break-out discussion groups” met on the second day of the IMS Workshop. Two of these groups focused on “Policy” and “Technical” issues, alternatives, and recommendations to be included in this report. A third group evaluated the feasibility of submitting a Category 4 Community Assistance Program (CAP) grant proposal to the FGDC and GeoConnections.

6.1 Policy

Several policies were discussed and recommendations were made in the Policy Break-out Group at the IMS Workshop. The groups agreed on the need for free and open data sharing, “new” and “value-added” data that should be made accessible if it is not available elsewhere and no restrictions exist on the data. In order to facilitate free and open sharing, the Common Data Needs TWG needs to create metadata guidelines for the TWGs and assist the groups in producing the metadata. This will be implemented in such a manner that anyone working in the study has access to all data. PIAG would have access to everything that has metadata (and/or has been reviewed). Any agencies, outside contractors, academics, and members of the general public external to the study can access any data for which metadata exists when it is produced by the study, but must go to the original owner if the data was not produced by the study. Free and open sharing could be complicated by the existence of sensitive data. Several types of data were identified that could be sensitive, and included modeled output (e.g. erosion lines, flood limits, and property values), climate change scenarios and water levels, marina data (e.g. competition), exact locations of variable and threatened species, intakes and outfalls, and detailed imagery.

Proprietary data issues were also discussed. The group decided that 1) any requests for licensed data should be directed to the original owner, 2) the IJC cannot assume ownership of licensed data, and 3) the Common Data Needs TWG should track licensing agreements. For new or value-added data, future ownership will be transferred to a willing recipient who would ultimately be responsible for the storage, archiving, and maintenance of the dataset. The new data owner should be carefully chosen to avoid transferring ownership to an agency unwilling or unable to make the data freely available. Licenses for commercially sensitive data should be noted in the metadata. The.

Data security and liability were also both addressed. It was agreed that data liability should be covered in the metadata. For all information products, a disclaimer is necessary. As specified in the Study Directives, the Board needs to approve all information products before they can be released to the public. [How this Directive will

be applied in practice, and to whom (if anyone) authority will be delegated needs to be resolved.] In terms of data security, the levels of data access must be defined, and specified within the metadata. Systems and mechanisms must be in place to ensure the security of the data. Overall, security for any licensed or proprietary data is the responsibility of the CDN TWG.

The issue of bilinguality of metadata and data was also addressed. The IMS team believed that the IJC policy of producing Study documentation in both English and French, and supporting translation costs, must be adhered to for metadata as well. The rationale for this requirement was that evaluation of the Study at its most basic level – the models that address Performance Indicators (PIs) under the “levels and flows” associated with different regulatory alternatives – requires the ability to evaluate the metadata associated with model inputs and outputs. Given that the majority of these models have a spatial component, it seemed prudent to provide geospatial metadata in both English and French. [Line item costs associated with translation of metadata are included in the evaluation of data discovery alternatives (below).]

While the Policy Break-out Group at the IMS Workshop felt that providing for bilingual metadata was justified, they did not support translation of the attribute information of the data sets themselves. Assuming that overviews of every model and supporting technical documentation will be translated according to IJC policy, the IMS team did not feel that the cost of translating the regional databases themselves was warranted. Given the above assumption, full evaluation of the modeling approach as well as inputs and outputs would be possible without translation of the databases. Moreover, the IJC should not bear costs of making every “new or value-added” geospatial dataset (including both PI-model inputs and outputs) immediately useful to the public, but only those costs necessary for full evaluation of the model approach. Translating metadata fulfills this latter requirement.

6.2 Technical

This group agreed that before technical needs could be addressed, interoperability standards needed to be specified. Also, GIS and database guidelines must be conformed to by all TWGs. To assist TWGs in compliance, it would be helpful to create very specific metadata instructions and a set of Frequently Asked Questions for technical issues. Metadata support for the TWGs would also help ensure that standards are met. The group anticipated that bilingual metadata would need to be produced (i.e., one English and one French version).

The group agreed that once standards were created, the technical needs of the project at different points in its life cycle could be addressed. They defined three main phases for

the project that affect technical requirements. The first phase is the current stage, which is infancy. At present, FTP and email are serving as the distribution mechanisms and distribution is driven by need/demand. There is a small volume of data and minimal use requirements, which keep the cost low.

The next phase is growth, which is expected to begin in the summer and fall of FY2002, and extend into FY2004. This phase will be characterized by increased volume of data to manage and an increased interest in data by the TWGs. The FTP site established as a temporary solution for the TWGs (currently managed by Ian Gillespie at CCIW) has accommodated much of the Study's data access and distribution needs to this point. However, the increased demand for data storage volume, speed of access, and site maintenance is expected to soon exceed this solution's capacity. Data discovery needs will increase among the TWGs, but not yet for the public. On-line interactive data browsing and mapping would be helpful in addressing many user and TWG needs. During the growth phase, security issues will be clarified. In the growth phase it might become attractive to utilize regional resources such as the IT divisions in Environment Canada Ontario Region and Environment Canada Quebec Region.

The last phase is maturity and includes contains everything after FY2004. At this point, most geospatial data and associated metadata should be complete and housed in a repository. The dissemination system should be operational. The amount of money necessary to run the system will increase with capacity and functionality. At this point the system will be driven by user needs. Mirrored sites may be needed for data distribution, after discovery through GLIN or GeoConnections.

6.3 FGDC (USGS) Cooperative Agreements Program (CAP) Grant Opportunity

Category 4 of the 2002 National Spatial Data Infrastructure (NSDI) Cooperative Agreements Program (CAP) provides an opportunity to acquire additional resources for implementing an information management strategy for the Study. The Joint U.S. and Canadian Spatial Data Infrastructure Project funds projects implementing and demonstrating the ability to address sound community decision-making through the collaborative use, maintenance and sharing of geospatial data over a common geography. FGDC and GeoConnections are collaborating to sponsor one project for the 2002 CAP. The award potential for this program is \$75,000 US provided by FGDC and \$100,000 Canadian provided by GeoConnections. This proposal requires 100% in-kind matching funds to be provided by the U.S. and Canadian partners respectively.

Prior to the Workshop, discussions with FGDC and GeoConnections provided some guidance for responding to the RFP. The group identified a preliminary set of public and

private partners, with the IJC serving as the lead organization for the proposed project, with Roger Gauthier and Wendy Leger as POCs, serving in their capacity as Common Data Needs TWG Co-chairs. Other public sector partners would include the U.S. Army Corps of Engineers, Environment Canada, Ontario MNR, and may include other provincial and state agencies. Pangaea Information Technologies and Great Lakes Commission would constitute the principle U.S. partners. CJS Consulting and Baird & Associates would constitute the principle Canadian partners. [Note: This organizational structure was changed in the actual grant proposal to have the ACE Detroit District and Environment Canada serve as the national Leads, with the IJC serving as a partner to both the US and Canadian groups. See Appendix IV for a summary of the proposed Project.]

Funds from this award would be matched by FY2002 funds already allocated for information management and would allow for greater attention to be given to the design and implementation of an information management system for the Study. Through the thorough documentation of the development process, made more feasible by the additional funding, the Lake Ontario and St. Lawrence River Study and its information management strategy and system design could serve as a model for other bi-national studies, particularly within the Great Lakes region. It should be noted that public outreach and the publicity activities are specified as components of any project receiving CAP funding. The group viewed the CAP project as an excellent compliment to the IM implementation activities recommended here, the PIAG's mission in particular, and the objectives of the Study in general.

The project would require a commitment to producing several specific data layers: these include geodetic control, cadastral, hydrography, elevation (topographic and bathymetric), political boundaries, transportation, ortho-imagery, and shoreline. The group proposed to have metadata stored at and accessed through the Great Lakes metadata clearinghouse (GLINDA) on the Great Lakes Information Network (GLIN). It would be necessary to develop a QA/QC team for review prior to metadata release.

These actions would meet FGDC goals (and our recommendations to the Study Board, below) by promoting, developing, and provide for training on metadata for the CDN and all other LOSLR TWGs. A guidelines manual would be created for reference purposes, and be made available to other groups. The goals of PIAG in promoting access to data and metadata would also be met, as well as the mid and long-term goals of data discovery.

If the grant were to be awarded to proposed participants, it has implications for data management, too. Some decisions would revolve around how the CDN is going to handle data management for the study and possibly beyond the study. Possibilities

include a single server, or a distributed network – consisting of servers outside firewalls which are maintained and updated by several agencies, such as LIO, the NYS GIS Clearinghouse, Quebec, Environment Canada, or US ACE, etc. A plan of action was proposed for use on the proposal for data, and consists of dedicating 2-3 servers to the data sharing effort and developing an interface page on the IJC project website which is basically an index of data, including FTPs of maps and provides options for download. The next steps to be taken include database development and connectivity for web mapping services (WMS) and web features services (WFS).

Questions that were still open at the end of the meeting were whether enough time existed to apply for the grant, how equipment purchases could be funded through this grant, and whether the proposal would mesh with the IMS Alternatives and Options (below) selected the Study Board. It was decided that flexibility in operations could help manage the CAP timeline and compliance with the Study Board desires, and also that there are other funding opportunities.

6.4 Study Properties and Needs

The IMS Workshop illuminated several new “Study Properties” and “Study Needs”, as well as clarifying or underlining those identified during the Needs Assessment Process (see section 3.3):

Study Properties (and Related Facts/Perceptions)

- Data discovery opportunities among (and even within) TWGs are presently limited.
- While communication and information/data transfer via email and FTP presently meet Study needs, this solution will likely be insufficient within 6 months.
- Standards-based metadata would be essential to the Study if: 1) Study datasets are going to be made public, 2) the value and longevity of Study datasets are to be preserved, and 3) automated discovery and evaluation tools need to be implemented in the Study.
- A fully-featured system that provides for data discovery, evaluation, and access for Study Participants has much potential to reduce redundancy and provide for a greater degree of integration among the modeling efforts with individual resource sectors.
- On-line, interactive mapping tools would aid in data evaluation for both Study Participants and the Public.
- On-line, interactive mapping tools would enhance transparency and public participation.

- The IJC does not wish to serve in a data stewardship or distributor capacity after the Study ends.
- For some datasets, data owners need to be identified by the end of the Study.
- New data owners likely will provide for the maintenance and distribution of Study datasets only for those political areas with which they are associated.
- A regionally-distributed storage, maintenance, and access system will likely preserve the value and increase the use of Study data.

Study Needs

- Metadata production is a present need.
- Standards that promote uniformity and interoperability need to be selected, communicated, promoted, and supported.
- CDN assistance with metadata creation for other TWGs.
- Support for compliance with standards (e.g., FGDC-1998 for metadata).
- Data discovery, evaluation, and access for Study Participants is a present need.
- Data discovery, evaluation, and access for the Public is a future need, but well within the life of the study.
- Free and open data sharing policy.
- Address sensitive datasets in policy and system implementation.
- Include data disclaimers and use restrictions in the metadata.
- Ensure system reliability and security.
- Data/information review process to have IJC, Board, or designated party confirm appropriateness for publication (see Annex 4c of the Plan of Study)

7.0 PRIMARY ALTERNATIVES AND OPTIONS (& ESSENTIAL SUPPORTING POLICIES AND COSTS)

7.1 Data Discovery

Data discovery and its associated mechanisms and functions provide the means by which information about the existence of data can be obtained. Current data discovery mechanisms require that some form of metadata, data about data, be compiled for each dataset and typically be made available in some organized fashion. The identified needs for data discovery and evaluation come from within the Study, in promoting Study-wide coordination of data, and from the associated desires to actively promote transparency in the Study and encourage public involvement. Geospatial data discovery has grown to become strongly associated with the establishment of spatial data infrastructures (SDI) in federal governments. The United States National Spatial Data Infrastructure (NSDI) and the Canadian Geospatial Data Infrastructure (CGDI) are the spatial data infrastructures which have implemented networks of data discovery clearinghouse nodes. More details of the SDI clearinghouse structure are presented later in this section of the report. A graphical depiction of geospatial data discovery using the SDI approach is presented in Figure 7.1.1 (below).

At the heart of any data discovery mechanism is the metadata from which information can be retrieved about the data. Data discovery is only as effective as the metadata generated for data is complete in content and quality. Metadata standards guide organizations in the creation of “complete” metadata by specifying format and content, and have been created by the FGDC and the International Organization for Standardization (ISO). Given the importance of metadata, not only in data discovery mechanisms but also in generally promoting clarity in data development, many organizations are committing resources to ensure metadata generation for all geospatial data and compliance with metadata standards. The Common Data Needs TWG has identified the need to support a single metadata standard and has committed to the FGDC-1998 metadata standard as that to which all technical working groups should comply. However, currently metadata does not exist for some of the data being used and produced by the study, complicating the ability to discover and cultivate data being used in Study activities.

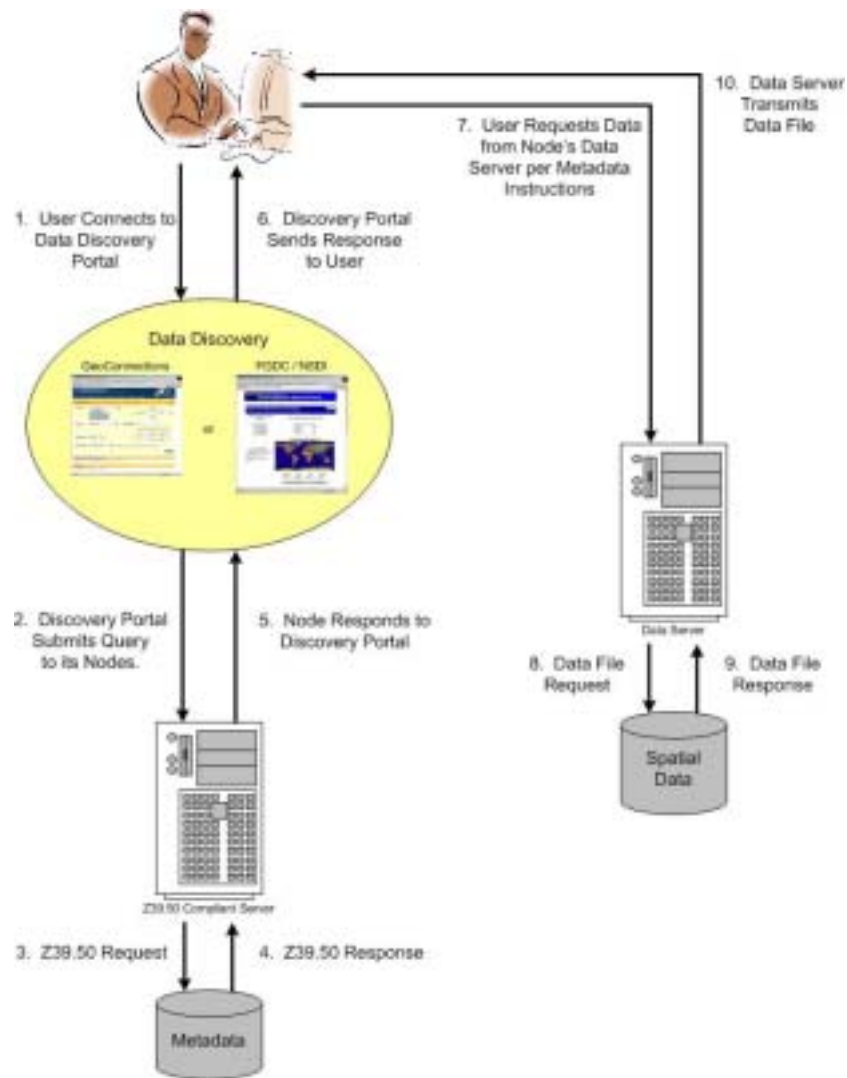


Figure 7.1.1 – Geospatial Data Discovery through the SDI

Study Participants (comprised of all the TWGs, the PIAG, and the Study Board) need the ability to explore within-Study data availability and to learn of the work of other study participants to meet the study mandate to “minimize data redundancy”. Potential benefits of a comprehensive data discovery mechanism include greater transparency in the system and a promotion of open dialog between resource sectors being evaluated. Cross TWG discussion and increased awareness of available (and planned) data sets will promote a more holistic approach to studying the impacts of water regulation schemes and generally creates greater understanding of the inter-related interests and concerns of various resource sectors associated with the Lake Ontario and St Lawrence system.

The public's capacity to discover and evaluate study data, modeling approaches, and results will promote greater public involvement in the Study. Again, the IJC and the LOSLR Study Board have expressed their commitment to public involvement and can, through implementing a data discovery mechanism, promote greater transparency in the Study methods. However, premature release of information and subsequent liability and security concerns make any sharing of information beyond summary information an important consideration. It is therefore necessary to establish clear rules and procedures for developing and publishing information in a public forum. While data discovery is a form of (summary) data distribution, the data being distributed (metadata) hold no significant liability or security concerns that warrant any reluctance in sharing the information. Data discovery simply shares information about the data and processing at a technical level within the Study.

7.1.1 Alternatives

Four alternatives have been identified for addressing the needs for data discovery. Each successive alternative addresses the need for metadata and data discovery to an increasing degree. Multiple alternatives can be conducted concurrently, and can therefore most effectively address different needs at various stages in the life of the Study. Additional options have also been identified that could be implemented in conjunction with the last two alternatives. Details for the alternatives, options, and evaluating criteria are detailed in the paragraphs below.

7.1.1.1 Status Quo

Data discovery performed in the LOSLR Study is currently a function of gleaning information from documents detailing the Study organization and work plans and/or by "word-of-mouth." In some cases limited information about data being used in the Study is included in such documents; however, the currency and completeness of such information is usually poor. The usability of existing documentation to learn about the data being used and generated in the Study is poor. The information contained in the documents would have to be reviewed in its near entirety to generate a comprehensive list of data; even then the data list would be only as current as when the source text was generated. Details concerning any specific data layers would not be available in the status quo alternative without first identifying the data owner and corresponding with them to inquire about metadata. Even if the data owner can be identified and contacted, the existence of metadata is not required in this alternative. The status quo option would require no additional funding or policy considerations; in essence, this alternative is already implemented.

7.1.1.2 Generating a Data List

The second alternative identified to address the need for data discovery involves generating a tabular list of all data used or generated by the Study. The tabular list would include general information about the data, its use and information about its ownership, maintenance, and distribution. The list would be distributed only to Study Participants. Because the information about the data is not compliant to metadata standards, it is not fit for public consumption. This alternative addresses the immediate need for inter-TWG data awareness in a limited manner, but does nothing to promote transparency and openness of the Study for the public.

Because the data contained in a table format is not parseable with standard metadata parsing engines, the functionality of searching the metadata table is limited to the searching functionality of the application through which the table is displayed. The most significant functionality that the data table alternative fails to support which is supported by other alternatives is the ability to search for data using geographic coordinates as part of the query.

The amount of effort and coordination necessary to generate the table and distribute it to the Study participants is, relative to the following two alternatives, minimal. Hence, this alternative is feasible in the near short-term. Because the information is not parsed or indexed, searching could only be done through basic text string searches on the distributed list. Other than a directive for TWG cooperation to provide brief metadata (name, scale/resolution, extent, data, etc), no additional policy considerations would be required in implementing this alternative. A full list of data being used and generated by the Study could be generated within ~2 months of initiating this alternative. Compiling the information necessary to create a comprehensive data list would require a coordination effort by the Common Data Needs TWG or independent contractor and time spent by Study participants involved in data development and analysis to provide for a comprehensive list. The basis for this list could be the list of dataset titles, grouped by TWG, which is presented in Appendix II.

7.1.1.3 Metadata Catalog

A catalog of standard compliant metadata would provide complete information about the data being used and generated by the LOSLR Study. A collection of metadata files, which can be made available over the Study website, represents a comprehensive list of information about data used in or produced by the Study. When generated in accordance with the FGDC standard, metadata files will include all information required to learn about (i.e., discover and evaluate) the particulars of data. Information regarding the party

responsible for data distribution is essential in order to acquire a copy. Information regarding the parties responsible for data maintenance is useful in providing feedback if errors are found, or for technical questions if such should arise. Furthermore, the information contained in the metadata would provide more detail in how the data is being used and/or generated by the Study. This alternative is the first to be fit for public consumption, addressing the need for public involvement and transparency in the Study process. It is important to note that a catalog available on the study website requires the data discoverer to visit in the Study website, thus improving visibility of the Study.

This alternative requires additional funding and a Study-wide commitment for the development and coordination of standard-compliant metadata. Standard-compliant metadata provides a common set of terminology and definitions to document data and allows an organization to maintain the investment made in collecting or generating geospatial data. Primary elements (text sections) of FGDC compliant metadata include: Identification, Data Quality, Spatial Data Organization, Spatial Reference, Entity and Attribute, Distribution, Metadata Reference, Citation, Time Period and Contact Information. These common elements and any specific data elements contained within allow users to determine things like the availability, fitness of use and accessibility of datasets. Commitment to standard compliant metadata will require some kind of metadata review process be implemented. Once reviewed, we expect that most Study metadata will require little or no maintenance through the Study's duration. The Study's dedication to metadata has strong implications on the functionality of a data discovery mechanism. Additional options may be selected to further assist in the metadata development and coordination. These options are listed in Section 7.1.2.

Policies in support of publicly discoverable metadata would need to be addressed in this alternative. All data and information produced by the Study should be made discoverable for the Study Participants and the public-at-large through a standard metadata documentation and collection procedure. Data used as model inputs that are *not* produced by the Study, and are readily available elsewhere, only would need to be cited appropriately in Study documentation. Regardless of the specific data or information restrictions, all metadata should be made accessible. In addition to providing for inter-TWG information discovery, such documentation and procedures will improve the visibility and transparency of the Study for the public, thereby promoting its overall credibility.

The consistency in metadata content and quality are crucial to the successful implementation of a data discovery structure. The development of data and information by any Study participant or contractor should be considered incomplete without compliance to metadata content and quality standards. A standard clause should be

included in all contracts related to data and information development, stating that required metadata is to meet all Study approved content and quality standards (see Appendix III for example).

7.1.1.4 Participation in SDI

The fourth and final alternative identified to address the need for data discovery is the participation in the spatial data infrastructures of the United States and Canada. The NSDI and the CGDI are both networks of metadata providers that use a standard search protocol to allow access to metadata through a single data discovery portal. Participation in the clearinghouse networks requires FGDC- or ISO-compliant metadata and a Z39.50 compliant server. Metadata is parsed and indexed when loaded onto the clearinghouse node, facilitating fully functional search capability. Once loaded, the clearinghouse is notified and the metadata is made searchable through the primary data discovery portal. In the case of CGDI, portal services could be accessed from a hyperlink within any website utilizing the Web API developed by GeoConnections.

Because participation in the SDI network requires the implementation of a Z39.50 compliant server, the Study would most efficiently utilize resources by submitting metadata to an agency or organization who has already implemented a clearinghouse node server. At present, the Great Lakes Commission (GLC) is establishing the Great Lakes Information Network Data Access (GLINDA) Clearinghouse, which is to be a clearinghouse node for the Great Lakes region. Participation in GLINDA or other metadata clearinghouse node such as the GeoConnections Discovery Portal would require minimal support by the Study, as the individual metadata files comprising the catalogue would simply be uploaded into the existing node. Once loaded onto a clearinghouse node, metadata would not require reloading unless updated to reflect changes to a dataset or contact information.

An international directory of SDI networks connects the nodes of different clearinghouses to create a world-wide network of metadata clearinghouse nodes. Discovery of the Study data in the SDI alternative can occur from multiple data discovery portals and nodes. This mechanism for data discovery increases the exposure of the Study and has the potential to attract the interest of more individuals than just those who would have otherwise known of or found the Study website.

The cost associated with this alternative would be similar to that of the metadata catalog alternative above, as the primary cost is related to the creation of metadata and support functions. While some support for the SDI node may be appropriate (requested or required), the additional expense would be minimal. As with the metadata catalog option

above, the four additional metadata options are applicable for the SDI participation alternative. Policies that support metadata creation, review, and uploading to a SDI clearinghouse, should be included as part of implementing this alternative. These policies would include adoption of the FGDC 1998 metadata content standard, ANSI Z39.50 compliance for server(s) holding the metadata catalog, and promotion of these standards at the contractual level.

7.1.2 Additional Options

Metadata Review Team:

To ensure compliance with metadata standards and consistent application of those standards, a metadata review team could be established for the purpose of conducting quality assurance and quality control on metadata as it is generated by the TWGs. The metadata review team could provide guidance in the development of metadata. [See description of “watch-dog” group formed for the Red River Basin Decision Information Network, Section 5.1.]

Metadata Coordinator:

A dedicated staff person could be responsible for ensuring metadata compliance to standards and consistency in metadata generated for the Study. As a short-term assignment, the metadata coordinator’s primary functions would be completed within the next fiscal year. Specific functions of the metadata coordinator could potentially include: coordination of all metadata training, providing assistance in metadata development, ensuring completeness of metadata produced, confirm compliance with FGDC 1998 metadata standards.

Metadata Workshop:

A metadata workshop held for all study participants involved in the production of metadata could provide for the necessary training and coordination to facilitate creation of standard compliant metadata. Training in metadata generation software could provide a jump-start to the metadata creation process, and reduce the time spent by a Metadata Review Team and/or Metadata Coordinator.

On-line Metadata Development Assistance:

On-line development assistance could help TWGs that are generating metadata through simple text instructions and easy to understand manuals. This mechanism could also include a mechanism to direct specific questions to an identified metadata expert (e.g., the Metadata Coordinator), who would be required to provide timely assistance. This option illustrates another means by which the creation of metadata can be facilitated to ensure fully compliant metadata for data layers.

7.1.3 Evaluation of Alternatives

Below are listed the primary criteria by which the above alternatives have been evaluated. A summary of the evaluation can be found in Figure 7.1.2.

Currency:

Currency in terms of data discovery is principally a concern of completeness. It addresses the question “Is the information about the data complete and up-to-date?” The Status Quo alternative involving the use of existing Study documents to glean information about data being used and generated in the Study lacks currency. The data defined within the planning documents available to all Study participants via the Study website is incomplete and only as up-to-date as the documents themselves. The three other alternatives would provide up-to-date information as they coordinate the identification of all data and provide the most updated information about those data.

Ease of Discovery by Study Participants:

The capacity for study participants to discover data being used and generated by other TWGs is consistent with the Study mandate of minimizing redundancy in data development. Inter-TWG data coordination is required to make the most efficient use of Study resources and to provide the highest level of integration among the resource-specific assessments. Therefore, facilitating data discovery within the Study fosters a more efficient use of Study resources, and an improved result. As the organization of information about data is the principle measure of ease of discovery, the Status Quo alternative fails to provide the necessary ease of discovery. The other three alternatives provide for the required organization of information about data to constitute ease of discovery.

Ease of Discovery by the Public:

The ability for the public to learn about data being used and generated by the Study is meant to address the need for transparency in the Study process. By providing a mechanism by which the public is able to learn of the data used in the Study, public involvement and acceptance of the Study results will be enhanced. Because information about data is published for public consumption only when it is standard compliant metadata, the Status Quo and Data Table alternatives fail to provide ease of discovery by the public. The Data Catalog alternative, while an acceptable collection of compliant metadata, would only be available to the public as a single document and would lack the search functionality necessary to constitute ease of discovery. The SDI alternative provides full access and search functionality through the SDI data discovery portal making this the alternative that provides the easiest public access, even to those not familiar with the Study.

Fit for Public Consumption (i.e., metadata standard):

Unless compliant with FGDC or ISO metadata standards the information about data is not appropriate to present to the public. Only completely compliant metadata will serve the public's need for data discovery and the Study's interest to include the public in the data discovery process. Because information about data is published for public consumption only when it is standard compliant metadata, the status quo and data table alternatives fail to provide ease of discovery by the public. The Data Catalog and SDI alternatives both require fully compliant metadata and are therefore fit for public consumption.

Comprehensiveness of Metadata:

Comprehensiveness of metadata refers to the completeness of the metadata in terms of metadata attributes. The Status Quo consists of no metadata, and the data table alternative consists of a very limited amount of information describing the data. Both the Data Catalog and the Participation in the SDI alternatives involve fully compliant (i.e., complete) metadata.

Organization of Metadata:

Only the Status Quo fails to organize information about the metadata; all other alternatives bring together a list of all the data and information about the data.

Fully Searchable:

An important function of any metadata discovery mechanism is the ability to search for information or specific characteristics about the data. The efficiency, flexibility, and thoroughness of the search function is directly related to the successful ability to and ease by which one can use the data discovery mechanism. While all the alternatives, at a minimum, involve digital information in one form or another that is searchable by text string, the SDI is the only alternative in which a search for geospatial data can accommodate geospatial (i.e., by x- and y-coordinates), categorical, and keyword searching.

Increased Exposure to the Study:

Unique to the alternative of participating in the SDI, data discovery portals not affiliated with the Study could be used to discover data pertaining to Study activities. In this case, the metadata provided to the user by the data discovery portal will contain information about the Study and who to contact for more information. In such circumstances, an individual interested in data for the region would be made aware of the Study and forwarded to the Study website by the metadata discovered through an unrelated source. This positive externality of implementing a networked data discovery mechanism promotes greater public involvement and awareness of the Study.

		Implementation Alternatives			
		Status Quo	Table	Catalog	SDI
Evaluation Criteria	Currency	▲	★	★	★
	Ease of Discovery (Study Participants)	●	★	★	★
	Ease of Discovery (Public)	●	●	■	★
	Fit for Public (i.e., metadata standard)	●	●	★	★
	Comprehensiveness of Metadata	●	▲	★	★
	Organization of Metadata	●	★	★	★
	Fully Searchable	●	▲	■	★
	Increased Exposure to Study	●	●	●	★
	Time to Delivery	★	■	●	●
	Cost	★	■	▲	●




Figure 7.1.2 - Evaluation of Data Discovery Alternatives

7.1.4 Costs

Data Discovery Budget Assumptions and Justification

- Unless noted otherwise, rates are calculated at \$475 per day. This is a blended average rate in which agency staff completes 75% of work at \$300 per day and the remainder by contractors at \$1000 per day.
- All costs are for labor, unless otherwise noted.
- All money is in US dollars.
- Estimated number of Study datasets is 200.
- Translation cost for each metadata file is \$124.44US.
- Amounts do not reflect yearly increase of salary and overhead.
- There are no costs associated with implementing the status quo alternative.

- Creation of metadata coincides with the completion of the data development and costs of metadata creation are proportionally distributed across the years of the Study in a 40%, 35%, 20% and 5% distribution scheme.
- Costs associated with metadata development may be reduced in whole or in part by selecting Options 2, 3, 4 or a combination thereof.
- Support of the SDI Node may be optional.

	Task	FY2002	FY2003	FY2004	FY2005	Study Total	Per Year After Study
Alternative 1: Status Quo	None	\$0	\$0	\$0	\$0	\$0	\$0
	<i>Yearly Total</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>
Alternative 2: Data List	Development	\$4,750	\$0	\$0	\$0	\$4,750	\$0
	<i>Yearly Total</i>	<i>\$4,750</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$4,750</i>	<i>\$0</i>
Alternative 3: Metadata Catalog	Metadata Development	\$11,875	\$10,391	\$5,938	\$1,484	\$29,688	\$0
	Translation	\$9,956	\$8,711	\$4,978	\$1,244	\$24,889	\$0
	<i>Yearly Total</i>	<i>\$21,831</i>	<i>\$19,102</i>	<i>\$10,915</i>	<i>\$2,729</i>	<i>\$54,577</i>	<i>\$0</i>
Alternative 4: SDI Participation	Metadata Development	\$11,875	\$10,391	\$5,938	\$1,484	\$29,688	\$0
	Translation	\$9,956	\$8,711	\$4,978	\$1,244	\$24,889	\$0
	Support SDI Node	\$2,000	\$0	\$0	\$0	\$2,000	\$0
	<i>Yearly Total</i>	<i>\$23,831</i>	<i>\$19,102</i>	<i>\$10,915</i>	<i>\$2,729</i>	<i>\$56,577</i>	<i>\$0</i>
Option 1: Metadata Review Team	Study Participant Time	\$5,700	\$4,988	\$2,850	\$713	\$14,250	\$0
	<i>Yearly Total</i>	<i>\$5,700</i>	<i>\$4,988</i>	<i>\$2,850</i>	<i>\$713</i>	<i>\$14,250</i>	<i>\$0</i>
Option 2: Metadata Coordinator	Agency Staff Salary	\$40,000	\$40,000	\$20,000	\$10,000	\$110,000	\$0
	<i>Yearly Total</i>	<i>\$40,000</i>	<i>\$40,000</i>	<i>\$20,000</i>	<i>\$10,000</i>	<i>\$110,000</i>	<i>\$0</i>
Option 3: Metadata Workshop	Participants	\$0	\$0	\$0	\$0	\$0	\$0
	Training Material	\$500	\$0	\$0	\$0	\$500	\$0
	<i>Yearly Total</i>	<i>\$500</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$500</i>	<i>\$0</i>
Option 4: Online Metadata Development Assistance	Implementation	\$4,275	\$0	\$0	\$0	\$4,275	\$0
	<i>Yearly Total</i>	<i>\$4,275</i>	<i>\$0</i>	<i>\$0</i>	<i>\$0</i>	<i>\$4,275</i>	<i>\$0</i>

7.1.5 Recommendations

Alternative 2 is recommended as a short-term solution to the need for data discovery to occur within the Study. In order for study participants to be able to learn of the data being used and generated by other study participant with enough time to allow for integration of data into analysis, a mechanism for data discovery needs to be implemented soon. For the purpose of the study, distribution of a list inventorying data and providing limited details is appropriate to meet the immediate need. The most important element of this alternative is the short amount of time needed to generate and distribute the information.

Alternative 4 is the long-term recommendation for the Study. Requiring the development of metadata standard compliant metadata, participation in the SDI initiative would provide a great amount of exposure with limited development effort given the utilization of existing resources such as GLINDA or GeoConnections Discovery Portal for implementing the metadata clearinghouse node. For the purpose of data discovery, loading of the metadata onto the clearinghouse node involves parsing and indexing which allows for greater search functionality. Information about the study's data would be stored on the node be available from any metadata clearinghouse portal, thereby increasing the potential exposure of the study to the public. This alternative is consistent with recent initiatives of both governments related to the coordination of geospatial data. By utilizing existing resources and providing fully compliant metadata, the study would be serving to promote the SDI initiatives supported by both the United States and Canadian governments and would at the same time be implementing a fully searchable data discovery mechanism to anyone with interest in the region or relevant data topic.

Options 2,3, and 4 are recommended in support of data discovery. These include: the creation of a staff position dedicated to the coordination of metadata for the Study, conducting a metadata workshop, and providing online metadata development assistance. The three options recommended support the creation of quality metadata, a crucial component of the data discovery process. A single metadata coordinator position would be responsible for ensuring that metadata generated by study participants were consistent and fully compliant to the FGDC metadata standard. The person assigned to this position would be accessible by study participants involved in the creation of metadata and could provide any necessary technical assistance in completing fully compliant metadata. A metadata workshop is another option chosen to support the process of metadata creation. As many organizations have only recently begun the process of generating standard compliant metadata, the expertise required to effectively comply with metadata standards is limited. A one and a half day workshop instructing participants on how to generate fully compliant metadata would improve the overall effectiveness in the study's metadata

creation tasks. The final option is a provision for online metadata development assistance. This option would include any software package or other user driven metadata tutorial that would assist study participants with simple, straightforward questions concerning metadata creation.

Our recommendations are to implement:

- Alternative 2: Creation of Data List (with brief metadata; a short-term solution)
- Alternative 4: SDI Participation
- Option 2: Metadata Coordinator
- Option 3: Metadata Workshop
- Option 4: Online Metadata Development Assistance

Total estimated cost of implementing the recommended alternative and options is \$73,356US in FY2002 and \$176,101.50 thru FY2005.

Policies essential in the implementation of the recommended alternatives and options are:

- The development of data and information by any Study participant or contractor should be considered incomplete without compliance to metadata content and quality standards (FGDC-1998).
- A data abstract, for use in data discovery, should be submitted with metadata.
- A data citation should be submitted with metadata
- A standard clause should be included in all contracts related to data and information development, stating that required metadata is to meet all Study approved content and quality standards.
- All metadata should be made available in both English and French. Translation of the datasets themselves is not required.

7.2 Data Storage, Access, and Distribution

A coordinated approach to data access is, at a minimum, essential for TWGs to complete their responsibilities in an efficient manner. Beyond facilitating the work being done within independent TWGs, the approach to data access has other implications for how effectively the Study is able to minimize data redundancy and ensure consistency between datasets and related analyses. The need for a system to facilitate data storage and access was repeatedly expressed in the Needs Assessment process. A data storage and access strategy has further implications on the extendibility of a system in accommodating the application of technologies such as the implementation of web services. Prior to determining how the Study will provide for data storage, maintenance, access and distribution, a clear understanding of the Study's commitment to facilitating

long-term sustainability and public accessibility to data and systems developed and utilized by the Study is needed.

Data being produced or significantly enhanced through the course of the Study will be the property of the IJC in most cases. The responsibility for ensuring the proper maintenance and presentation of data, while held by the IJC, will likely be assigned to the TWG associated with the data. In this scenario, data ownership is held by the IJC and the responsibility of storage, maintenance, access and distribution are assigned to TWGs (or TWG members) serving in a data stewardship capacity. At the fruition of the Study, the IJC may no longer desire or be able to continue in its role as data distributor, and thus be willing to forego its role as data owner. The responsibilities of TWGs will be discontinued, preventing them from serving as data stewards. Therefore, under the current scheme the sustainability of data is probably limited to the life of the Study, after which point it would exist in data archives. The value of the data for use in evaluation of Study results, in facilitating further studies, or for a variety of other uses (some undoubtedly unforeseen at present) continues well beyond the defined life study. Therefore, the need for accommodating long-term sustainability of data and systems should be addressed from the beginning of the information management strategy implementation.

Public accessibility of Study data has significant implications and is most efficiently addressed in concert with providing accessibility to all Study participants and related agencies. Access to data and information utilized and/or produced by the Study should be determined through a rules-based procedure considering the data's ownership, security, liability, licensing, privacy, and proprietary status and the relationship of the interested party to the Study. The following simple rules for access have come out of discussions and correspondence throughout the information management strategy development process.

- All primary Study participants (e.g., Study Board, PIAG, and TWG members) should be given access to all data and information utilized and/or produced by the Study, with the exception of data and information having special security, liability, privacy, licensing, or proprietary concerns.
- All other interested parties should be given access to any data and information which is considered new or having value added to it by activities of the Study, with the exception of data and information having special security, liability, privacy, licensing, or proprietary concerns.
- “New data or information” is defined as that which did not exist prior to Study activities and was generated from primary data collection procedures as a direct result of Study activities, i.e., model output or results.

- “Value-added data and information” is defined as that which has been significantly improved as a result of Study activities in either its content or usability, *and* cannot be readily accessed elsewhere.
- Data and information deemed “sensitive” (i.e., possessing special security, liability, privacy, licensing, or proprietary concerns), should be systematically tracked by the Common Data Needs TWG. The CDN TWG should track all licensing agreements.

These guidelines can be applied in assigning rights and privileges in a coordinated data storage, maintenance, access and distribution strategy. Additional considerations in implementing a data access strategy will include specific liability and/or security concerns associated with the premature release of data to public scrutiny. Careful steps will need to be taken when datasets become mature and public accessibility is addressed. Products of the Study will need to be thoroughly reviewed and any disparate opinions among Study members regarding results or appropriate use should be addressed prior to the data being published. Disclaimers and appropriate use restrictions will need to be presented (e.g., in the metadata), to anyone who wishes to acquire data.

7.2.1 Alternatives

Four alternatives have been identified for addressing the needs for data storage, maintenance, access and distribution. While the implementation of multiple alternatives simultaneously was possible for data discovery, the alternatives here are much less compatible, with the possible exception being the temporary “implementation” of an extended status quo to accommodate the short-term needs of the Study during the development, testing, and final implementation of a better alternative. Additional options have also been identified that could be implemented with either of the two more functional alternatives. Details for the alternatives, options, and evaluating criteria are detailed in the paragraphs below.

7.2.1.1 Status Quo

The current data storage and access scheme implemented for the Study allows users (Study Participants) to store and access data in their local environments. While this typically would provide easy access to data by those who are connected to the local environment in which data is stored, access to data by other users normally requires the use of an FTP site, where data is uploaded by the source user and then removed by the destination user. Other mechanisms for data transfer involving various media (e.g., CDs, magnetic tapes, etc.) are also likely being used. The system for data distribution is largely uncoordinated and fails to facilitate data integrity, security, back-ups or archiving.

This system includes no active maintenance functionality for individual datasets: incremental changes to parts of a dataset could not be made, and only full replacement would be possible. While the FTP site being managed by Ian Gillespie at CCIW has accommodated much of the Study's data access and distribution needs to this point, the increased demand for data storage and site maintenance is expected to soon exceed the capacity of this temporary solution. Considerations for public accessibility of data and long-term sustainability of data and systems have not been addressed under the current strategy. No immediate additional costs are associated with continuing with the status quo; however, because the status quo FTP site was intended as a temporary solution, a decision to continue with this strategy will likely require that additional capacity be added in the near future as the demand for its use increases. No additional policy considerations are essential for the persistence of the status quo; however any considerations of public accessibility will require a coordinated Study policy, particularly in a less coordinated storage and access strategy.

7.2.1.2 Single Repository

The second alternative identified to address the need for a coordinated data storage, maintenance, access and distribution is the implementation of a single repository for Study data. The repository would exist as single FTP site to which users can be assigned rights and permissions according to their specific information needs. As a single location for all Study data, the repository would allow for much greater coordination of data distribution. Data integrity, security, back-up and archival would be facilitated in a single environment. The repository would be able to accommodate public access to data through providing limited access with read-only permissions or by implementing a webpage with hyperlinks to FTP downloadable files.

While more coordinated than the status quo, a single repository has limited potential for facilitating long-term data sustainability. Data owners and corresponding data stewards with the ability, interest and motivation to ensure long-term data sustainability are likely to be less willing to manage data in a single system (read: national and provincial concerns and legal issues). As with the preceding alternative, this system would preclude the possibility of active maintenance functionality for individual datasets.

This alternative, like the two following it, will require a flexible data security model to be implemented. Such an approach for granting rights and permissions has and can be easily implemented from a technical standpoint without significant effort. (From a policy standpoint, of course, this is not so straightforward.) In addition to managing file security, a common data transfer standard (e.g. SDTS) or de facto standards (e.g. shapefile or .e00) will be necessary to provide consistency across the study. The

additional costs associated with the implementation of a single data repository would include the expansion of additional storage volume on a system having ample bandwidth to accommodate the need for data transfer associated with data distribution.

7.2.1.3 Single Data Base Management System (DBMS)

The third alternative identified to address the need for a coordinated data management strategy involves the implementation of a single data storage, maintenance, access, and distribution system. Establishing a single system in which data is loaded and stored in a relational database environment will facilitate the full integration of data into a comprehensive system. A database system in which data is stored in a logical structure will allow for data to be integrated into other systems and accommodate the application of other technologies much more effectively than through using a file structure. The single location will facilitate data integrity, security, back-up and archiving. However, because long-term sustainability is dependant upon the willingness and ability of data owners and stewards to maintain datasets, as with the previous alternatives this one prohibits long-term sustainability by inhibiting regional ownership and stewardship. A single system could potentially alienate the regional partners who are removed from the system location. The long-term sustainability of the single system is tied to the motivation of a single maintainer to manage it beyond the life of the Study.

Policies to provide for appropriate public accessibility would need to be established under the single system alternative. Similar to the single repository, a flexible data security model and standards for data transfer would need to be implemented. Costs associated with the single system alternative include hardware, software, development, training, implementation, and maintenance. Additional options for enhancing functionality represent add-ons to the single system alternative. The three options are listed in Section 7.2.2.

7.2.1.4 IJC Distributed DBMS

The fourth alternative identified to address the need for a coordinated data management strategy involves the implementation of a data system similar to the single DBMS described above, but divided and managed by the respective national offices of the IJC in Ottawa and Washington DC. A dual system would be developed and maintained in a consistent and interoperable manner so as to support seamless data access across national jurisdictions. By committing to the development and maintenance of systems managing data for the LOSLR Study by national jurisdiction, the IJC would build an information management infrastructure to support the data management needs of the LOSLR Study, and potentially, future studies. This option offers direct control over almost every aspect

of systems development, implementation, and maintenance, without reliance on the coordinated effort of other agencies to form a functional information infrastructure. However, it does not take advantage of the exiting pool of available resources nor the long-term benefits associated with a more distributed, regional approach.

This alternative would require the Study Board's support to equip the IJC national offices with the necessary hardware, software and expertise required to develop, implement and maintain interoperable geodata management systems. Because this approach requires the development of IM support staff and resources, the cost associated with this dual system is substantially greater than the regionally distributed alternative, which takes advantage of the infrastructure and established knowledge base of other regional organizations. However, while the cost is associated directly with the LOSLR Study's IM system development, implementation, and maintenance, it could also be considered an investment for future studies and other IJC information management needs. So long as the IJC would choose to support and maintain scalable data management systems, future studies could take advantage of the infrastructure created by this alternative as well as the knowledge base established within the IJC as a result of the development and maintenance of the systems.

Insofar as the IJC commits to maintaining a dual data management system, the long-term sustainability of the *system* will be provided for. Certainly, for the duration of the study, the IJC has the necessary motivation to maintain the new and value-added *datasets* that form the basis for many of the Study's recommendations. But unless access to the datasets is necessary after conclusion of the Study (e.g., for continued public review, or for use in other IJC studies), the IJC would no longer have a direct need for or motivation to maintain the datasets. It will be important for the IJC to weigh this future need when deciding upon which alternative is most suitable: this system would be cost-effective only if other IJC studies can utilize it after FY2005. If future IJC needs would not be met, then the large investment associated with this nationally-distributed alternative would be outweighed by the advantages of less expensive alternatives, as will be discussed below.

7.2.1.5 Regionally Distributed DBMS

The fifth alternative identified to address the need for a coordinated data management strategy involves the implementation of a data system similar to the single system described above, but divided and managed at the regional level. The establishment of regional data management systems most effectively addresses the need for regional partners to ensure the longevity of data associated with the Study. As with data owners, regional system maintainers would need to be identified just as data owners would. This

data management model is the most progressive, and is endorsed and promoted by the public sector (FGDC, GeoConnections), private sector (CubeWerx, Inc.), and NGOs (OpenGIS Consortium).

The regionally distributed systems would be developed in a coordinated effort to ensure maximum consistency in system implementation and maintenance. As described above for the single system alternative, a regionally distributed system would take advantage of relational database environment to manage the structure of the data store. Again, this facilitates greater integration and connectivity to other systems, and can more easily accommodate other technologies such as web services. At a minimum, interoperability standards would be specified (and need to be adhered to!).

Preliminary evaluations of regional resources (introduced in Section 4.1), which could facilitate system development and ensure system reliability, have identified the following agencies and organizations. For the United States region of the Study area, the New York state options preliminary investigations revealed no system or organization with the capacity to provide for the information management needs of the Study. The GLC may soon secure non-Study funding to support the development of a database-driven information management system at the University of Michigan, which could serve as a major component of the LOSLR–US Region Information System. This system is being designed with the scalability to provide similar information management services for other studies within the Great Lakes region.

For the Ontario region of the Study area, LIO and Binational.net have been identified as potential regional partners. LIO is a fully functional information management system running on a database environment and is able to support WMS and is developing support for other OWS. Binational.net is being developed by EC-Ontario Region and the US EPA in order to accommodate binational programs conducted between those agencies. While a possible option, Binational.net is less developed than LIO and may introduce bureaucratic challenges in implementing a flexible system to accommodate the Study's information management needs. For the Quebec region of the Study area, ECQR has been identified as the forerunner for implementing an information management system to accommodate the Study's needs. Presently in the implementation stage of a database driven information system behind their firewall, knowledge gained from this experience would facilitate the development of the system the Study outside of the firewall, as well as provide a knowledge base for development of the GLC DBMS.

While all three systems in a regionally distributed information system will have separate administration, development should concentrate on consistency to ensure a common approach to data storage, maintenance, access, and distribution. In addition to addressing

seamless system development and implementation, *data* held in the systems will be clipped to a common boundary and/or need to be made seamless in order to facilitate the overall consistency of the Study data. The costs associated with establishing a regionally distributed information management system for the Study include hardware, software, development and implementation. System development for the regionally distributed information management system will require additional time in comparison to the single system development to accommodate the additional coordination of effort and system implementation. Options associated with this alternative are identical to those listed for the single system alternative but with additional considerations due to the distributed nature of this alternative. Because the Study data is distributed across three servers, it will be necessary to implement middleware on each of the regional systems to allow a single web service to utilize all three stores of data. Implementation costs associated with the middleware would approximately double those of the middleware option in the single system alternative.

7.2.1.6 TWG Distributed DBMS

A final alternative that should be considered to address the need for coordinated geospatial data management involves implementing data systems similar to the “Single DBMS” described above, but with components distributed among TWGs. This approach has several advantages. However, these are confined largely to activities that will take place during the duration of the Study.

In the TWG Distributed DBMS alternative, maintenance responsibilities would be determined by whether: 1) the system resources were acquired by the TWG specifically for its work for the Study, or 2) the resources were on loan from a participating agency. In the scenario in which resources are on loan from a participating agency, the system resource management responsibilities likely with that agency, while data management responsibilities exist with the TWG in whatever arrangement they deem reasonable. In the scenario in which resources are acquired by the TWG specifically for implementation of the system, the TWG is responsible for both maintaining the system and data, and for finding a suitable location for their system. At present, the latter scenario is likely for all TWGs with the exception of Hydrologic & Hydraulic, and Coastal.

A geodata management system distributed among TWGs would place the data and system in relatively close association with the data developers and initial data users. As such, reliable access and control over the data has the potential to increase the overall motivation required for system upkeep during the Study. Moreover, because the system and geodata would be managed by that data’s primary user-group, data currency and integrity should remain up-to-date.

At the conclusion of the Study the data and system resources acquired by the Study would be transferred to the IJC; resources on loan from a participating agency would be released and data relying on those resources would be transferred to the IJC, or identified data owners whenever possible. Because this approach includes datasets that encompass international and provincial boundaries, unlike the “Regionally Distributed DBMS” alternative, securing data owners with the motivation to provide for database maintenance beyond the Study’s terminus could prove problematic.

Data transfers between TWGs will rely heavily on the ability for other Study Participants to connect to and access data from the distributed network of servers. Similarly, for Study Participants or the public to simultaneously access multiple geospatial databases that are distributed among TWG servers for interactive data viewing or map-making (WMS-related activities), interoperability is essential. In this alternative, the development and exact system configurations of the TWG systems have the potential to vary considerably, based on the specific demand and resources available. Thus, inasmuch as possible, each system component (typically one per TWG) should be standardized for consistency across the Study. Insofar as the systems are standardized, interoperability and the potential for providing connective features such as web services would be promoted. To this end, a specific system design and configuration (i.e., “standard build”) including network operating systems, middleware, software, and database structures should be developed and adopted.

This alternative would require the Study Board’s support through the allocation of funding required to implement a large network of distributed systems, one for each individual TWG. Without specifying design criteria here, it should be stated that each TWG would need a server for data accessibility (both inter-TWG and to the Public), and a workstation with software capable of geospatial data management and production purposes. As with the previous two alternatives, a database (DBMS) rather than file system approach is recommended. Cost estimates for this alternative (below) reflect these specifications.

Cost estimates do not reflect the unique problems associated with the Lake Ontario – Upper St. Lawrence River Coastal Data Server (CDS). Although the exact status of implementation is unknown at the time of this report, a file system approach was recently proposed. Although the CDS could be used within this Alternative, hardware and software would need to be standardized to the greatest extent possible with the other TWG servers, including a database management system (DBMS) rather than a file system structure. This type of dilemma, (starting work v. waiting for an IM strategy and specifications) points to the need of IMS efforts addressed earlier in the Study life cycle.

7.2.2 Additional Options

Interactive Data Viewing and Map Making (i.e., WMS Capability):

OGC Web Services is a set of web-based services developed by the OGC to facilitate the open source application of mapping (WMS) and GIS functionality (WFS) over the Internet. Implementation of OGC Web Services will require a connection to the database in which data layers are stored. Costs associated with implementation of this option include program development to customize the services for use with the Study's data.

Support Proprietary Internet Mapping Services:

More highly developed than the OGC Web Services, proprietary Internet mapping services can provide more robust functionality, i.e., involving geospatial operations such as overlaying or proximity analysis. The largest disadvantage with implementing proprietary Internet map services is the cost associated with purchase and licensing.

Implementation of a Database Middleware:

Middleware is used to connect applications to database management systems (DBMS) environment. The implementation of middleware in the single system alternative would allow for remote access to data layers stored on the system from desktop GIS applications or web services located on different servers.

7.2.3 Evaluation of Alternatives

Graphical depictions of the flows of information between modeling groups (each associated with a particular TWG) and DBMS servers, and how alternatives 3, 4, 5, and 6 differ in this regard are presented as Figures 7.2.1 (a-d).

Below are listed the primary criteria by which the above alternatives have been evaluated. A summary of the evaluation can be found in Figure 7.2.2.

Interactive Data Viewing and Mapping (i.e., WMS Capability):

Data viewing is the ability to simply look at the spatial data on-screen through a web mapping viewer. A web mapping service (WMS) produces maps from data located in a structured data store as images in an Internet-enabled environment. The structure of the data storage system dictates the ability of that system to accommodate WMS implementation. The Status Quo and Single Repository alternatives lack the necessary structure in the file system to accommodate WMS. All the other alternatives provide the structure required and therefore the capability to support WMS. The ability of the two distributed alternatives to support WMS relies heavily on interoperability and the

consistent implementation of standards in their disparate systems. Systems implemented using different technologies hinder the interoperability required in providing cohesive data viewing and mapping.

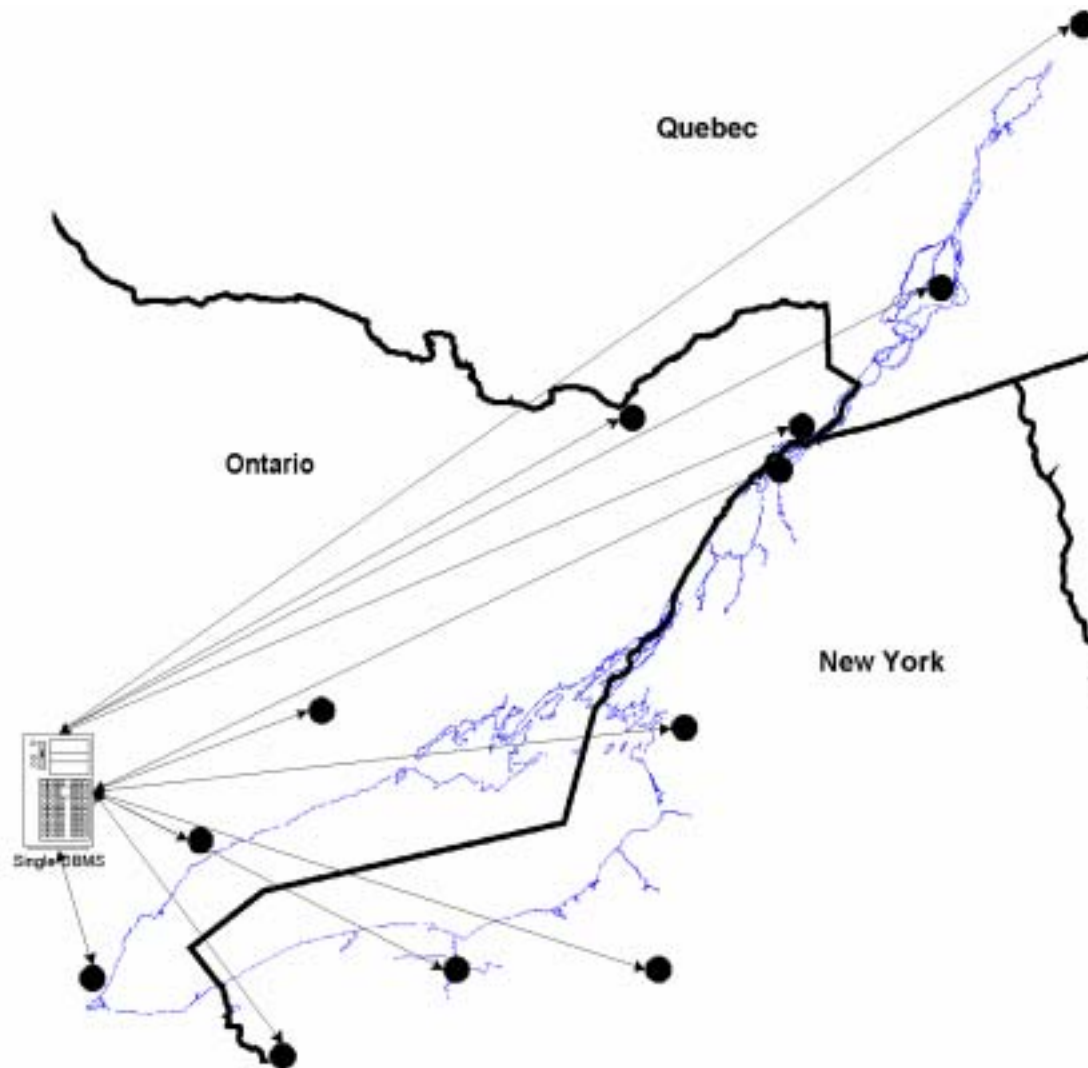


Figure 7.2.1(a) - Flow of information between modeling groups and Single DBMS server in Alternative 3.

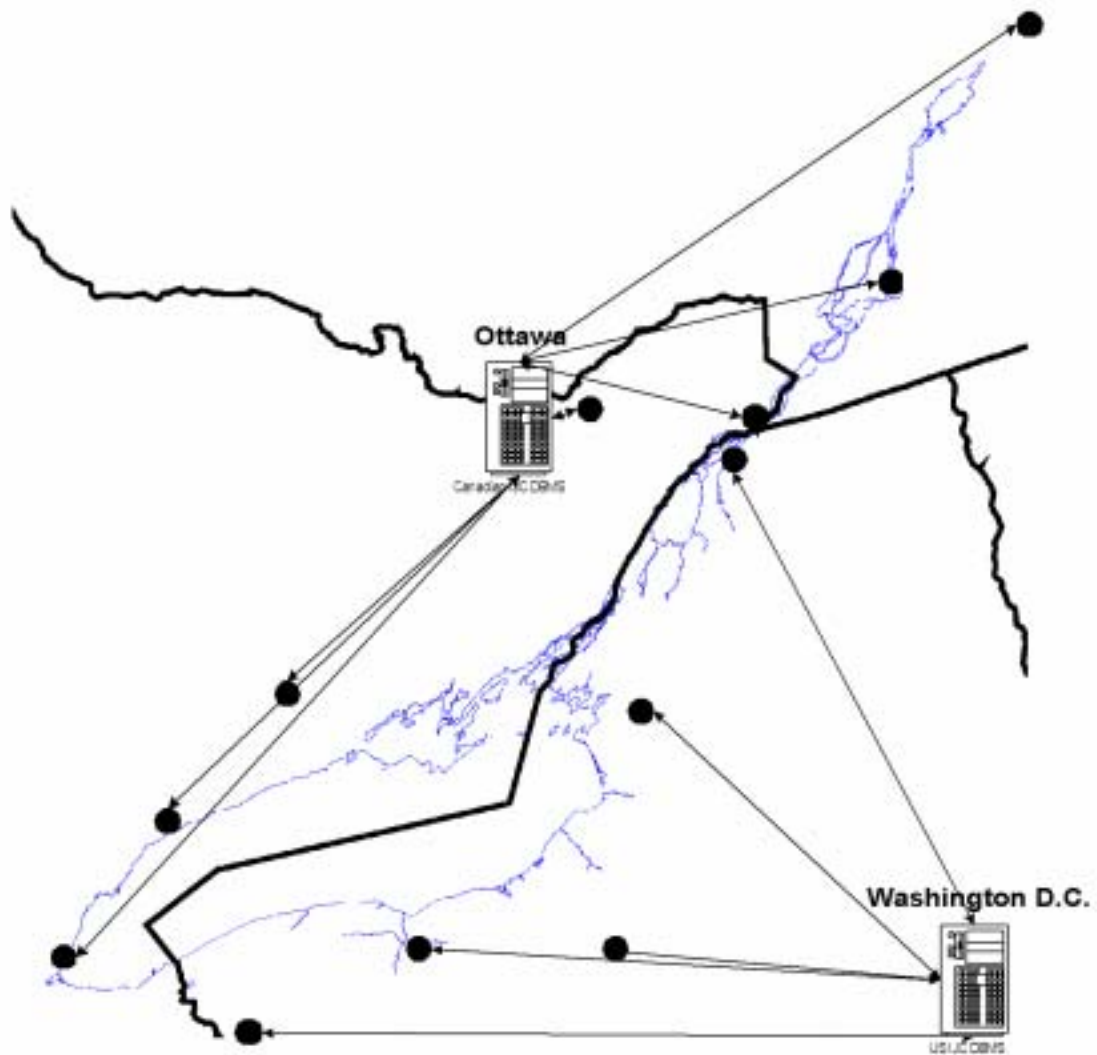


Figure 7.2.1(b) - Flow of information between modeling groups and IJC Distributed DBMS servers in Alternative 4.

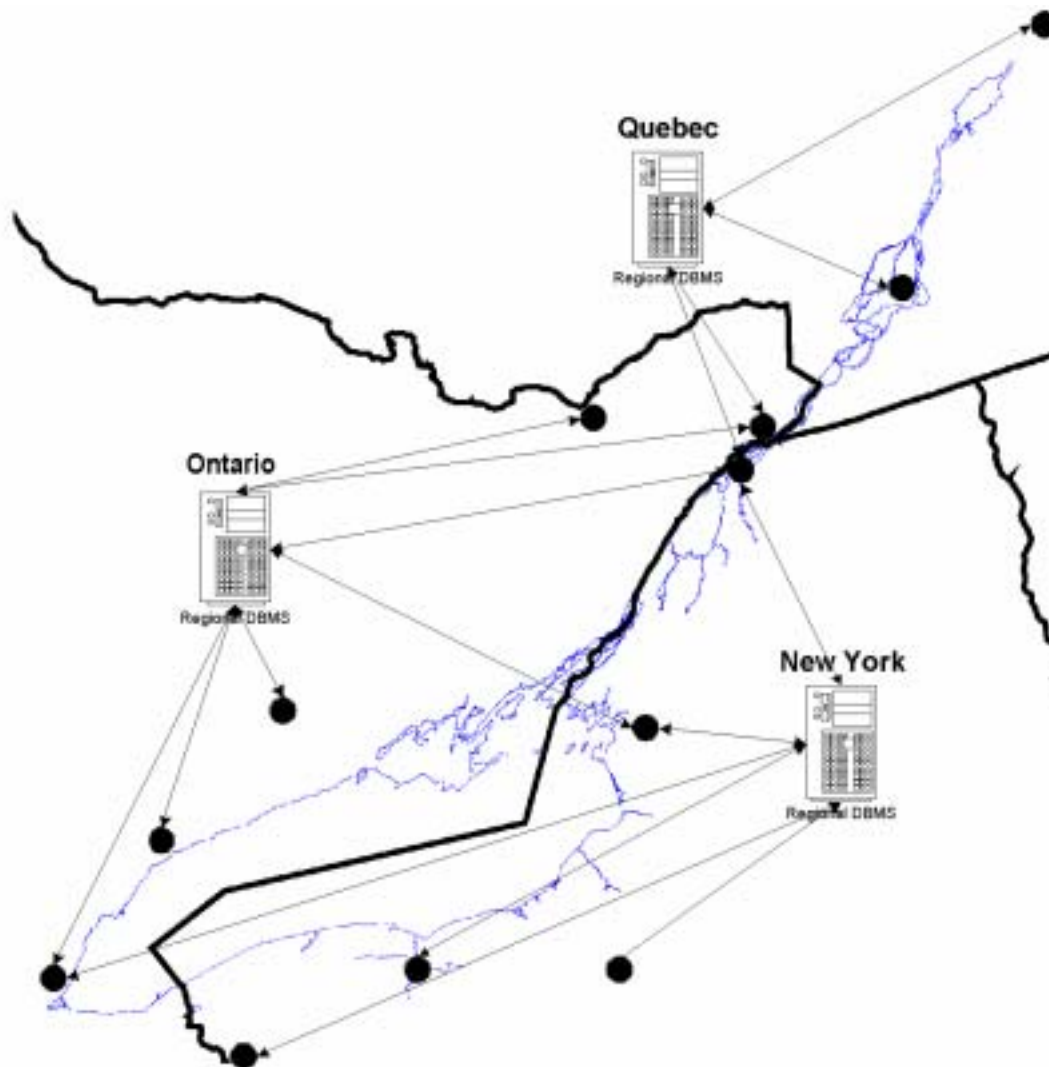


Figure 7.2.1(c) - Flow of information between modeling groups and Regionally Distributed DBMS servers in Alternative 5.

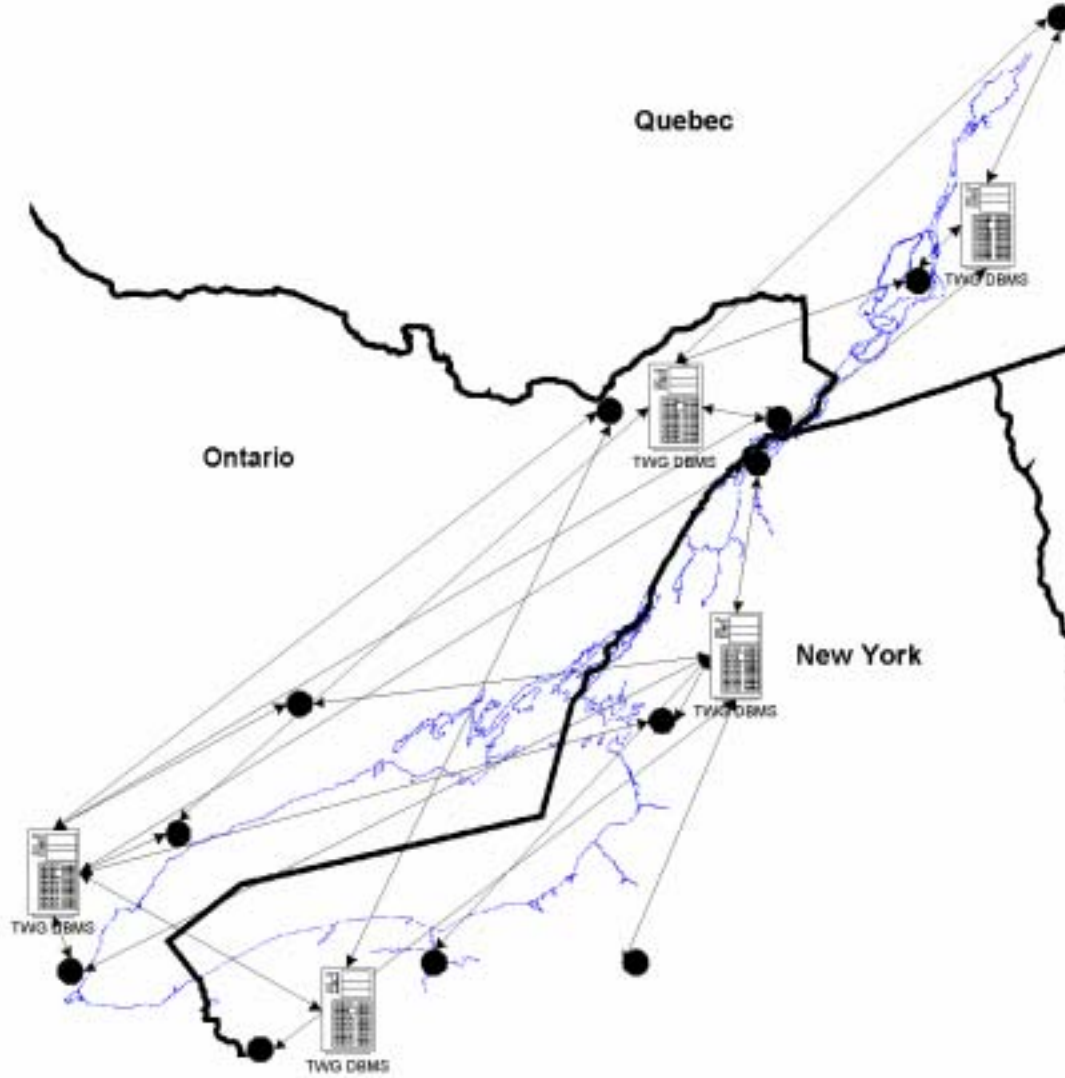


Figure 7.2.1(d) - Flow of information between modeling groups and TWG Distributed DBMS servers in Alternative 6.

Capacity for WFS:

The web feature service (WFS) can support more robust functionality in geospatial services than can WMS, but the technology is less well developed. A WFS allows for greater flexibility in developing web-based custom processes with GIS data. As a burgeoning technology, WFS considerations should be considered if future information management and system development is desired. The structure of the data storage system dictates the ability of that system to accommodate WFS implementation. The Status Quo and Single Repository alternatives lack the necessary structure in the file system to accommodate WFS. All the other alternatives provide the structure required and therefore the capability to support WFS. As with WMS, the ability of the two distributed alternatives to support WFS relies heavily on interoperability and the consistent implementation of standards.

Potential for Long-term Sustainability of Data:

“Long-term sustainability” of data is directly related to the ability to identify new owners willing to take responsibility for data beyond the life of the study. It is unlikely that data owners can be identified who are willing and able to ensure long-term sustainability of data and data access within a system housed in a region different than their own. This does not preclude a regional authority’s acceptance of a single Study-wide DBMS; however, an agency’s or organization’s familiarity with systems managed within another agency or organization of their own region increases the likelihood of identifying appropriate data owners. Potential for regional ownership and stewardship indicates an ability to identify a committed, regional level agency or organization. The activities (and data) associated with regional agencies are controlled by the policies dictated by regional authorities. Thus, the feasibility of implementing a regional ownership and stewardship arrangement is directly tied to the acceptance of such a system by these regional authorities. Acceptance by the regional authorities is more likely to occur in an environment familiar to them in which some manner of control can be exercised. The Regionally Distributed DBMS is the only alternative that completely supports the potential for regional ownership/stewardship and therefore long-term sustainability of the data. The Status Quo alternative lacks the organization required of any ownership/stewardship scheme that would support long-term sustainability of the data. The other three alternatives could, but are unlikely to, support an ownership/stewardship scheme that provides for long-term sustainability of the data.

Consistency of the System:

The consistency of the system indicates the Study’s ability to manage data and provide for access in a way that is seamless in method and content. The consistency of the system is not an evaluation of the seamlessness of the data, but instead it evaluates the amount of work it would take to display the seamless data in an appropriate manner.

The Status Quo alternative lacks the coordinated organization to be considered consistent. The distributed alternatives have the potential to support consistency across the network of disparate systems. The Regionally Distributed DBMS approach has better potential for system consistency than the TWG Distributed DBMS approach because of the fewer total number of systems to coordinate. This distinction highlights the problem created by trying to coordinate disparate systems that are technically complex. The evaluation of consistency in the distributed DBMS alternatives reflects the effort needed to ensure the appropriate level of interoperability between the various components of the system. The IJC Distributed DBMS approach has better potential for system consistency than the Regionally Distributed DBMS approach for two reasons: in the former approach, there would be one fewer DBMS to coordinate, and the two systems would be controlled by a single authority (i.e., the IJC). The Single Repository and Single DBMS alternatives have the greatest degree of consistency because they are solitary and isolated approaches, not relying on the coordination technical complexities. The ultimate test of consistency in an information management system for the LOSLR Study, particularly in the regionally distributed approach, would be to evaluate the results of an information request encompassing the New York – Ontario – Quebec border area. Failure to return a consistent result would indicate potentially significant discrepancies in the system design and/or content that could affect the effectiveness of the Study in general.

Ease of Accessibility by Study Participants:

The ease of accessibility by Study Participants is necessary to ensure the efficient use of study resources and is one of the most critical of the evaluation criteria. Without a reliable system for ensuring Study Participants' access to information necessary to complete their responsibilities, the information management strategy would be a failure. While the Status Quo allows for access to data primarily on a by-request basis, the other alternatives all provide a mechanism for Study Participants to transfer data within the Study, though the Single Repository alternative does so in a relatively cumbersome manner.

Ease of Accessibility by Public:

The Public's ease of accessibility has significant implications on the public's acceptance of the Study and the credibility of its results. Furthermore, the value of data utilized and generated by the Study may warrant its provision to the public if only as a service provided by the Study. While the Status Quo provides no mechanism by which the public can access to data, the other alternatives all provide some mechanism in support of public accessibility to Study data. Again, the Single Repository alternative does so in a relatively cumbersome manner.

Foster Study Transparency and Facilitate Public Involvement:

Directly related to the ease of access by the public, fostering Study transparency is intended to allow the public to become more involved in the Study and knowledgeable of its methods and results. Study transparency is necessary throughout the life of the Study, and possibly for some time thereafter, to ensure the credibility of the Study. Study transparency is promoted across the alternatives as the range of functionality and extensibility of the various alternatives increase. Because of the TWG Distributed DBMS alternative's complexity, it is evaluated as having lower potential in promoting transparency in the Study than the IJC and Regionally Distributed DBMS alternatives.

Provide Model for Other Organizations and Studies:

The IJC LOSLR Study has the opportunity to provide an example for other Studies of comparable scope to use as a template for the design and implementation of an information management system that utilizes information technology to advance the efficiency, effectiveness, and public participation in a public sector study. In essence, this criterion is an aggregate evaluation of the potential support of the functionality and extensibility of the alternatives.

Long-term Sustainability of the System after the Study:

Similar to the concern of long-term sustainability of data, sustainability of the Study information management system will be determined by the willingness of the system maintainer(s) to keep the system up after the conclusion of the Study. The utilization of the Study by the public will warrant the necessity of maintaining the system beyond the life of the Study, as will the investment of public funds in the system. The Status Quo and TWG Distributed DBMS alternatives fail to coordinate data or functionality at any level greater than the TWG of the Study. The Single DBMS is the most coordinated alternative, but is evaluated as poorly supporting long-term sustainability after the Study because of the limited potential for a politically and technically feasible scheme for long-term ownership/stewardship of the data given the orientation and complexity of the system. The Single Repository alternative is evaluated more favorably because the long-term sustainability of the system would require minimal work given the relative simplicity of supporting an FTP site versus a DBMS. The Regionally Distributed DBMS alternative is evaluated most favorably because it directly accounts for and incorporates the interest of organizations and agencies required to support long-term sustainability. Proper evaluation of the IJC Distributed DBMS alternative depends on the IJC's commitment to maintain such systems in the long-term. Given the investment necessary for developing and implementing an IJC Distributed DBMS approach, it is likely that the IJC would plan on sustaining the system for some period of time after the completion of the LOSLR Study.

Potential for Study-wide Backup and Archiving:

As a function of the Study’s data management needs, data backup and archival is necessary to protect Study resources. Much of the data being used by the Study will not need to be made easily accessible after a particular phase is complete; however, a record must be maintained. The principle measure for evaluating this criterion is the whether an organized collection of the data exists. In the Status Quo alternative no organized collection exists and therefore the potential for backup and archiving is not supported. In all the other alternatives some form of backup and archival can be supported in either a single or distributed fashion.

		Implementation Alternatives					
		Status Quo	Single Repository	Single System	IJC Distributed System	Regionally Distributed System	TWG Distributed System
Evaluation Criteria	Data Viewing (i.e., WMS Capability)	●	●	★	★	★	★
	Capacity for WFS	●	●	★	★	★	★
	Potential for Long-term Sustainability of Data	●	▲	▲	▲	★	▲
	Consistency of System	●	★	★	★	■	▲
	Ease of Accessibility by Study Participants	▲	■	★	★	★	★
	Ease of Accessibility by Public	●	■	★	★	★	★
	Foster Study Transparency and Facilitate Public Involvement	●	▲	■	★	★	■
	Long-term Sustainability of the System after the Study	●	■	▲	■	★	●
	Potential for Study-wide Backup and Archival	●	★	★	★	★	★
	Provides Model for Other Organizations and Studies	●	▲	■	■	★	▲
	Time to Delivery	★	■	▲	●	▲	●
	Cost	★	■	▲	●	▲	●

Figure 7.2.2 - Evaluation of Storage, Maintenance, and Access Alternatives