Administration

Fostering Digitization Initiatives

In Saskatchewan

A Discussion Paper

Prepared by

The Digitization Working Group
With representation from

University of Saskatchewan Libraries and Archives
University of Regina Libraries and Archives
Saskatchewan Archives Board
Saskatchewan Legislative Library
Regina Public Library
Saskatoon Public Library
Provincial Library, Saskatchewan Learning

April 2003
EXECUTIVE SUMMARY

This paper proposes a methodology for Saskatchewan institutions to develop strategic alliances with one another to enhance their digital content creation capacity.

Many Saskatchewan organizations are developing digital content. Archives are creating digital objects from fragile historical documents and making them accessible on the Internet. Libraries of all types are digitizing popular information resources. Schools, colleges and universities are creating digital course materials for Internet-based education programs. Museums are creating digital objects to showcase their collections. Many cultural organizations and other information providers are also creating digital content.

This paper proposes that interested institutions form a strategic alliance to:
- share information about their individual digitization initiatives;
- discuss issues related to digitization projects -- digital rights management issues, metadata standards, hardware, software, and product design;
- develop a body of expertise, best practices, and standards;
- share two resource people who can become advisors on digitization;
- investigate joint licensing of digitization software/hardware;
- establish a funding pool for digitization projects; and
- represent the digitization interests of participants to vendors, potential funding sources, and government.

The following pages outline a draft model for developing a strategic alliance on digitization. We invite Saskatchewan institutions to review the model and give us feedback. We need to know if this type of initiative would meet your needs and have your support. As a draft, this model is expected to evolve. Your feedback will shape the outcome.

DIGITIZATION ADVISORY ALLIANCE DRAFT MODEL

VISION -- Working Draft

To develop a locus of knowledge and support that enables the public to find, locate, and utilize Saskatchewan knowledge and information created through co-operative Saskatchewan digitization initiatives.

VALUES -- Working Draft

Co-operation would be founded upon shared principles:

- Partnership
- Access
- Preservation
- Institutional learning
- Education
- Cultural development
- Promotion
- Standards

Partnership -- The value of strength in collaboration.

Access to and discovery of information -- Commitment to access to information.
**Institutional learning** -- Commitment to growing as institutions by learning new skills, and sharing our learning experiences with our partners.

**Education** -- Digital content is an important component of education: K-12, post-secondary education, and life long learning.

**Cultural development** -- Commitment to preserving cultural heritage and developing our cultural future via digital content creation.

**Promotion** -- Digital content is an effective way to promote our organizations and the province.

**Standards** -- Commitment to excellence in digitization by fostering adherence to international digitization standards.

**Preservation** -- Capturing Saskatchewan information.

**OBJECTIVES -- Working Draft**

1. **Build expertise**

Digitizing materials and building digital content is complex. Agencies must make decisions regarding copyright, digital rights management, metadata standards, technical issues (hardware, software), and product design. Information demonstrating the complexity of technical issues and standards is included in the appendices to this paper.

We propose finding a way to fund two individuals: an expert on copyright, digital rights management, and legal components of digitization; and, an expert on international metadata and digitization standards, as well as digitization hardware and software. These individuals could be shared among the alliance's partners, providing direction, support, and on-site assistance.

We also propose engaging alliance partners in regular meetings to facilitate sharing of expertise and experiences.

2. **Provide co-ordination and leadership**

We believe it would be useful to have a locus of provincial co-ordination and leadership to:

- Co-ordinate joint licensing of digitization products, if desired.
- Inventory digital content developed.
- Create a registry of Saskatchewan digitization projects with detail about the standards and policies they have used (i.e. best practices).
- Compile priority lists and wish lists for future digital content creation and for navigational tools.
- Co-ordinate collective decision making on matters such as standards.
- Co-ordinate regular meetings of the alliance partners.
- Identify, co-ordinate, and/or build partnerships around joint projects where desirable.

3. **Develop financial capacity**
Most institutions apply for external grant funds to finance digital content creation. Many granting agencies require applicants to contribute a certain dollar amount to the project. This process could be aided through the establishment of a provincial funding pool to be used for leveraging grant funds for digital content creation. The alliance could establish a pool and solicit contributions. The alliance could set up an impartial adjudication group to review applications.

4. Future proof digital collections

Many institutions are currently building digital content. With this content scattered in separate repositories across the web, it can be difficult to find. Some digital objects are stored in software repositories that may be proprietary and inaccessible.

"Future proofing" means utilizing international digitization standards so that digital objects created in isolation can, at any time in the future, be "virtually" integrated into broader digital collections, libraries, learning resources, teaching tools, and other future environments and formats.

Future proofing increases the value of digital objects because they can be dynamically integrated into broader collections and information products. Future proofing also increases the value of digital objects because it facilitates the building of a single search interface for accessing various digital collections and making them easily available to all users. International digitization standards are based on open-source computer programming, which means that they do not rely on proprietary fee-based software to access them, thus making them more affordable to operate and access into the future.

We propose sharing information on future proofing and international standards, so that institutions in Saskatchewan have a better understanding of their options when developing digital content. Information sharing will also enable institutions to make some collective decisions about technological standards, thus enabling future interoperability of their digital repositories.

5. Preserve Saskatchewan Information

Digital standards have not yet evolved to the point where they offer a permanent alternative to traditional methods of reformatting original documents. Digitization does, however, offer the ability to provide researchers with a virtual image that can be used instead of the fragile original, thus saving the original from further damage through frequent physical handling (although it may be noted that demand for physical handling of the original can often increase as a result of virtual publication).

Preservation is also a concern (for information managers) where solutions must be found for issues of technical obsolescence and migration for 'born-digital' documents, i.e. information originally created and stored in digital formats, and print documents that have been transferred to electronic form. A digitization advisory alliance could help institutions address these concerns by providing guidance on digital standards that offer the best long term accessibility options.

From these perspectives, digitization can be seen as a useful adjunct to paper preservation programs. A priority of a Saskatchewan digitization alliance should be to build understanding of the opportunities and limitations that digitization offers with regard to preservation, so that partner institutions can make wise preservation decisions.
STRUCTURE -- Working Draft

Structures for this initiative will be investigated if there is sufficient interest from potential partners. Structures would then be evolved with the input of the partners.

We suggest the following principles and priorities to guide structural development.

- **Locus of knowledge concept** -- The focus should be to build and co-ordinate a collaborative support network (instead of building administrative structures).

- **Expert staff** -- Expert staff, committed to the digitization alliance is necessary given the complexity of digitization. Funding would need to be sought for two staff persons, who could possibly be based in a partner organization.

- **Governance** -- An advisory committee, with representation from all partners, could be established for the alliance. The “home” for this advisory committee and staff would need to be determined in consultation with the communities of interest. The Multitype Library Board is a possible option because it has a role to co-ordinate co-operative province-wide initiatives for all types of information providers in Saskatchewan.

- **Funding pool** -- Financial contributions must be sought to develop a funding pool. Light administrative structures, managed through partner representatives, the parent/host organization, and staff members, are a possibility.

- **Provincial government support** -- Sustainable funding should be sought from the Government of Saskatchewan, including funding to hire staff support and for the funding pool. Government support is appropriate given the broad positive impact this initiative could have on the learning, library, heritage, archives, and cultural sectors (particularly in regard to technology-enhanced learning and online service development).

- **Additional financial support** -- Funding should be sought from other groups, such as the Canadian Council of Archives and other bodies identified by the communities of interest. Corporate donations should be sought to support a funding pool.

**REQUEST FOR FEEDBACK**

We believe that embracing digital technology will enable us to meet the future needs and expectations of our respective clients, and, that an alliance will enable us to undertake digitization projects more effectively.

We would like your feedback on this strategy. Is digitization important to you? Would this type of initiative meet your needs? Would this initiative have your support? Would you like to be involved? Would you like to attend a follow up meeting with other interested organizations?
APPENDIX

Issues Concerning Digitization Methods, Standards, Support Requirements, Copyright, and Digital Rights Management

This appendix was first written by Brett Waytuck, formerly of Provincial Library, following his attendance at the University of New Brunswick Electronic Text Centre’s Summer Institute. The Summer Institute was held in Fredericton between August 20-24, 2001 and was taught by David Seaman, director of the internationally recognized Electronic Text Centre at the University of Virginia. Revisions to the Appendix have been made by the Digitization Working Group. The Working Group would like to thank Mr. Waytuck for his digitization research, out of which this provincial strategy has evolved. The issues identified herein articulate the complexity of digitization and represent the type of expertise that could be collectively built and shared through a cooperative alliance.

A. Methods of Digitization

Standards

Markup Languages

The past several years have seen the development of internationally recognized standards for the preparation of text and images for delivery in various electronic formats.

The most important of these is Extensible Markup Language (XML). XML was derived from Standard Generalized Markup Language (SGML), an ISO standard in its own right. It is a set of rules for describing data and designing text formats that let people structure their own data. It uses a Document Type Definition (DTD) or an XML Schema to describe the data.

With XML, the data with different formats can be converted to a single format that can be read by many different types of applications. Industries can use XML to define platform-independent protocols for the exchange of data, allowing people to structure, store, retrieve, and display information the way they want it.

What makes XML so important is that it provides a means of marking up a text once and then, through the creation of front end style sheets, distributing the text through a variety of electronic media (XHTML/Web based, pdf, e-books, palm devices). Also, as XML is based on recognized standards there is the expectation that any new format or device will be compatible with a properly marked text -- eliminating the need for backwards conversion except through the creation of new style sheets.

Encoding Standards

The most pervasive and important encoding standards for the digitization of texts are the Text Encoding Initiative (TEI), Encoded Archival Description (EAD) and Dublin Core.

TEI was developed to aid in the conversion of humanities and social sciences texts and manuscripts to electronic formats, but other forms of information such as images and sound are also addressed. The TEI Guidelines describe an encoding scheme that can be expressed using a number of different formal languages. The first editions of the Guidelines used SGML; but the most recent edition can also be expressed in XML.
EAD was developed as a way to create electronic finding aids for archival collections. It is a set of rules for preserving the hierarchy and designating the intellectual and physical parts of archival finding aids to help search, display and exchange archives and manuscript collections. The EAD rules are written in the form of a SGML Document Type Definition (DTD), because archival description emphasizes intellectual structure and content more than bibliographic description, making SGML, and later XML, a more suitable transport syntax than MARC (MARC is the standard for bibliographic information in machine-readable format, used by libraries).

Dublin Core Metadata Element Set (DCMES) was designed as the descriptive metadata to support digital resource discovery. It has the similar functions as MARC to organize digitized data. Dublin Core provides a simple core set of description elements that can be used by normal users who are not familiar with cataloging rules for simple digital resource description. The core set of elements could be modified and extended by adding new elements according to the specific requirement of the project within the framework. This provides DCMES flexibility, interoperability and extensibility.

These encoding schemes provide a significant amount of internal flexibility. Individual projects and document classes can be evaluated for the level of internal markup required to effectively access the text, while at the same time insuring that the base electronic document conforms to a standard and will interact with other documents similarly encoded.

Adhering to the XML standards also allows for the creation (usually in conjunction with digital imaging of the source document) of an electronic archival copy of the document.

TEI and other XML encoding languages do not require, but do not preclude accompanying digital images (scanned or photographed) of the source material. Where these aid the viewer, or are required due to expectations of the end user, they can be linked to the encoded document for clarity. Where they are not required the encoding agency can provide a faithful reconstruction of the original text without the added expenses of imaging and electronically storing the original.

The process of using standards for digitizing and its advantages and disadvantages can be summed up as follows:

<table>
<thead>
<tr>
<th>Description</th>
<th>Digitizing texts via scanning or keyboarding and marking the electronic text with a standardized tags to aid in searching, retrieval and display.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Allows for “reading” of the document in a manner similar to the physical item. Allows for searching of the document(s) at a level controlled by the viewer and the encoding agency. Opportunity to clarify information at any level. Ability to easily convert source material into forms readable by any electronic text delivery device.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Time and costs involved in marking up entire text.</td>
</tr>
</tbody>
</table>

**Other Methods of Digitization**

While the XML family of markup languages and style sheets are increasingly recognized as library and industry standards for the encoding of texts, there are other methods for digitizing materials.

While it may be argued that adhering to the recognized standard is the preferred course of action this is not always practical based on a variety of considerations. These considerations may include the cost of the digitization versus available funds, intended
use of the encoded document, grant or funding requirements, compatibility with related materials, copyright restrictions and historical precedence.

A selection of common alternatives to TEI encoding are described below.

1. **Raw HTML / XHTML Encoding**

<table>
<thead>
<tr>
<th>Description</th>
<th>The framing of a text with HTML / XHTML codes for display on the Internet.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>None given the power of the XML standard markup languages and the ease of their post-production conversion to HTML / XHTML.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>The amount of work involved in creating these electronic documents is almost equivalent to that involved in creating an XML encoded document without the resulting benefits of standardization.</td>
</tr>
</tbody>
</table>

2. **Digital Imaging With No Text Encoding**

<table>
<thead>
<tr>
<th>Description</th>
<th>Creating &quot;photographic&quot; images of the document (via scanning or digital photography) for retrieval.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Provides the viewer with a faithful reproduction of the original document. Allows for &quot;reading&quot; of the document in a manner similar to the physical item.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Inability to search the documents in any manner. Readability dependant solely on the quality of the original document and the resulting scan/photograph. No opportunity to clarify unclear information. Images require large virtual storage space. Print quality dependent on image quality.</td>
</tr>
</tbody>
</table>

3. **Digital Imaging With Metatagging**

<table>
<thead>
<tr>
<th>Description</th>
<th>Creating “photographic” images of the document (via scanning or digital photography) and adding meta tags at a variety of levels to aid in retrieval.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Provides the viewer with a faithful reproduction of the original document. Can allow for “reading” of the document in a manner similar to the physical item. Allows for searching of the document(s) at a level controlled by the encoding agency.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Readability dependant solely on the quality of the original document and the resulting scan/photograph. Opportunity to clarify unclear information only at the macro level. Does not allow for natural language or keyword searching. Images require large storage space with limited text retrieval. Print quality dependent on image quality.</td>
</tr>
</tbody>
</table>

4. **Digital Imaging With pdf**

<table>
<thead>
<tr>
<th>Description</th>
<th>Scanning a source document into Adobe Acrobat’s software for Web reading and display.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Provides the viewer with a faithful reproduction of the original document. Can allow for “reading” of the document in a manner similar to the physical item. Allows for keyword searching within the document. Very good print results.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Proprietary software, not an open source standard. Readability dependant solely on the quality of the original document and the resulting scan/photograph. Cross document searching not perfected.</td>
</tr>
</tbody>
</table>
5. Text Entry into a Searchable Relational Database with Accompanying Digital Imaging

<table>
<thead>
<tr>
<th>Description</th>
<th>Creating “photographic” images of the document (via scanning or digital photography) and entering the full or partial text into a relational database for searching.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Provides the viewer with a faithful reproduction of the original document. Can allow for “reading” of the document in a manner similar to the physical item. Allows for searching of the document(s) at a level controlled by the viewer. Opportunity to clarify information at any level. Allows for natural language or keyword searching.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Readability dependant solely on the quality of the original document and the resulting scan/photograph. Images and database require large storage space. Relational database will not meet international standards. Difficulty in sharing source documents with other collections. Print quality dependent on image quality.</td>
</tr>
</tbody>
</table>

B. Digitization Support Requirements

Besides issues related to digitization standards, all projects have certain technical and personnel requirements. These requirements may vary depending on the type and size of the project at hand, but a general rule is that the larger the project the more equipment and personnel it will require.

Outlined below are some of the technical and support personnel requirements that may be utilized in a digitization project. This listing is not intended to be exhaustive and there has been no attempt to provide “industry standards” information. In relation to the descriptions of required personnel, it should be kept in mind that people often perform more than one role while working on a project.

Equipment

Computers and Software: The technical requirements can vary greatly dependent on the project at hand. Digital imaging requires access to software and computers capable of capturing, editing and displaying images separate from the text creation computers. If keyboarding is done locally, computers are required for the work. Software can also be purchased to aid in this but is not required.

Digital Cameras: Used for archival electronic imaging of materials. Also utilized for materials too fragile for flatbed or OCR scanning. Mounting systems, lighting and lenses are also required for more advanced treatment of materials.

Imaging Structures: Purchased or built equipment utilized in the digitization process. Requirements vary with the fragility and rareness of the material to be digitized. Can include such things as scanner extensions, camera mounts, lighting systems, book rests.

Scanners: Depending on the project, standard flatbed or OCR (optical character recognition) scanners may be required.

Servers: There must be someplace to both store the information that has been digitized and to make it available to the user community. Locally owned servers have the advantage of being under the direct control of the content creators. The public is not in as much danger of losing access to the digital information. The servers need to support the digitization processes used.
Personnel

Administrators: Responsible for establishing goals of the project. Coordinate funding, initiate grant proposals, secure copyright clearances, negotiate scope of projects, work with text authors and publishers, establish quality control objectives, market and promote collection.

Designers: Responsible for designing web pages, web display, e-book or other electronic media display.

Imagers: Production people responsible for scanning or taking digital photographs of materials. Must be able to work with equipment and software. Must be able to maintain established quality control guidelines. May need to have training in archival handling of materials.

Intellectual control personnel: Identify type of digitization to be done, identify technical requirements, establish and apply metadata requirements, design associated database requirements, establish and monitor quality control levels.

Keyboarders: Production people responsible for translating printed/written text to an electronic environment. Can be onsite or offshore. Accuracy rate of scanning with OCR technology at its best only meets minimum rate guaranteed by double keyboarding / computer comparison of texts. Keyboarders better at dealing with manuscript or unique type face materials. Keyboarding required for anything but “dirty” OCR scanning. Require training in whatever markup or tagging language is being used.

Systems people: Insure goals of project can be met with available technology. Coordinate equipment/software needs and compatibility, design associated databases and interfaces, work with intellectual control personnel to insure technical specifications are executable and met, insure storage capacity and file linking needs are met, insure accessibility.

C. Copyright and Digital Rights

As with all forms of publishing, copyright is a concern. The perception that copyright is more complex in a digital environment is only partially correct. Fundamental copyright rights and restrictions currently still operate in the digital environment. Within the context of this discussion paper it is not possible to itemize all of the copyright concerns related to the digitization of text and images. There are some general issues, however, which are appropriately discussed within this context.

What is often not understood is that digitizing a text or image and then posting it to the Internet, storing it in a computer, or transferring it to a CD-ROM is the equivalent of re-publishing the item.

Institutions that undertake digitization projects must recognize early in the process that they are undertaking a publishing initiative. Full discovery and securing of necessary rights is a basic requirement. Without this discovery and securing of rights, institutions can face the prospect of having the time, money and equipment they have invested in digitizing a document wasted.

Institutions must also be aware of the copyright and digital right responsibilities that digitization carries. While there is a public perception that what is available on the Internet is “copyright free” this is obviously not the case. Whatever digitization work a library completes has digital rights associated with it. Even if the copyright of the original
source has expired, the newly created digital resource is copyrighted. Institutions must ensure that it is clearly understood who holds the copyright for the newly created work, who will have responsibility for safeguarding the copyright, and what digital rights the users have when they use or purchase the digital resources.

If the material is placed freely on the web, the concerns regarding copyright may be minimal (although it is unlikely that the originating institution would want someone else to copy the digital files and distribute them as their own). If the material is created with revenue generation in mind, then copyright issues will be of greater concern – insuring that rights are protected, royalties paid and material is not copied illegally.

Pursuing this discovery and securing of rights can, however, have unintended positive benefits for both the institution and its community. By working within their community, institutions are more likely to be able to explain to their patrons the advantages of digitizing important materials; insure that materials created by and for the community remain freely accessible to that community; ensure that text and image electronic rights remain within the community or the province; ensure that knowledge management principles are consistently applied to digitization projects; and further demonstrate to stakeholders that the institution is an integral and living part of its community.