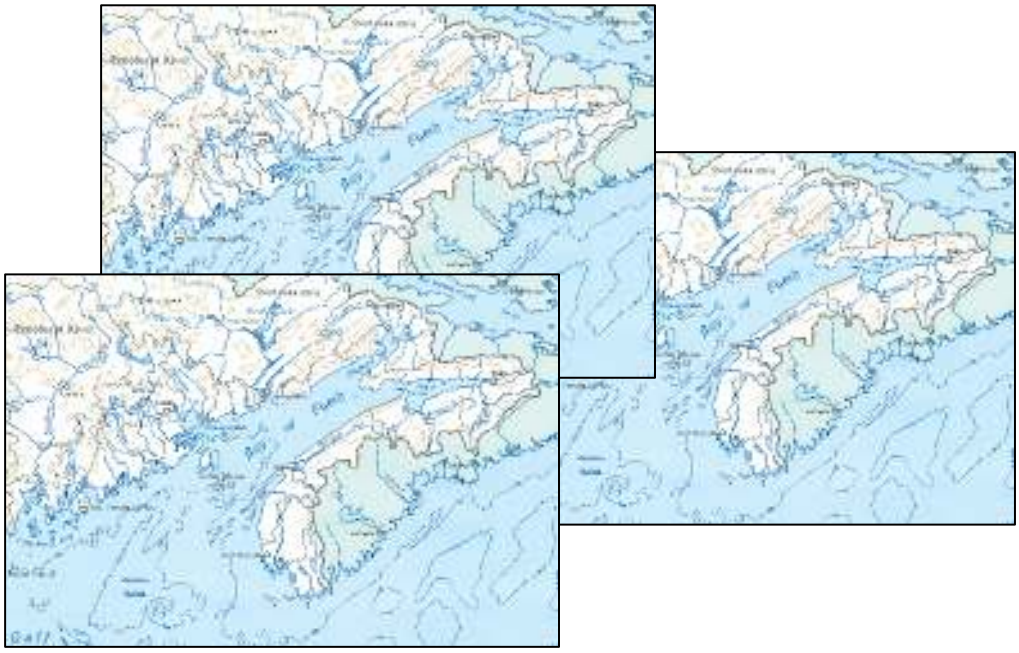


Developing an Information and Knowledge Repository for the Bay of Fundy

Final Report to BoFEP

Elaine G. Toms
Bertrum MacDonald
Peter Wells

December 2007

Introduction

Like many organizations and agencies, the wardens of the Bay of Fundy and Gulf of Maine are faced with an ever increasing volume of information and data that needs to be sifted through when making decisions about the care and management of the Bay and Gulf. The challenge has been in finding the information and/or data that are needed to support and facilitate decision making and problem solving. The challenge is exacerbated when one examines the range of 'wardens' – from policy makers to citizens and scientists, and the range of possible information sources and resources. This situation is not novel to this group, but extends to many other environments. A recent survey of 1000 managers in Canada and the US by Accenture Ltd. found that managers waste around two hours a day looking for information and much of what they find is of little value to them. Too often, wrong or inappropriate information is retrieved. Often the challenge is not that the information does not exist, but that it is difficult to find and access, among an ever increasing volume of available information.

Our goal in this project was to rethink the design of current tools for providing access to information about the Bay of Fundy. Like many organizations, BoFEP is driven by information use and knowledge creation. It is a valuable resource to the scientific and policy-making communities for ongoing environmental management of the Bay and the Gulf. To date much of BoFEP's information is contained in printed publication lists, bibliographies and proceedings. While indexes have been provided, in particular, to enrich access to the papers in printed proceedings, the contents of these papers and reports are still hard to find and awkward to access.

Objectives

The purpose of this project was to aggregate the documents that have been created and/or compiled by the Bay of Fundy Environmental Program (BoFEP) to date, and to create a web-based and accessable, full-text, digital repository of those documents. The intent was to provide Google-like access to the intellectual resources of BoFEP, and to enhance that access by enabling searching for documents by a variety of metadata as well as to enhance the browsability of the resources.

Because of the limited resources applied to the project, this was to be a demonstrational project – a working prototype with a small collection. Such a system could be used to leverage additional funds for a more sophisticated system compatible with existing Gulf of Maine/Bay of Fundy systems.

Background

Over the past few decades, access to information has evolved with changes in technologies from the traditional approach – using bibliographic databases – to extended systems that provide full text search engines, to augmented systems that provide even greater flexibility enabling search and browse capabilities. This range of systems has been limited to the 'bag of words' approach' to accessing the content of documents. The next stage in the evolution of content-rich systems will be the *enhanced* system that not only provides full text search engines enriched using text mining approaches, but also considers the contexts of use. Enabling the provision of information at the point of decision making is the ultimate goal.

An extended system is illustrated by the current Gulf of Maine Council on the Marine Environment (GOMC) web site (www.gulfofmaine.org), with the list of knowledge-base topics, lists of papers and their PDFs, and the link to the search engine Google. An augmented system is more like the web site maintained for the USA Chesapeake Bay Program (http://www.chesapeakebay.net/), where, for example, one can browse through a range of environmental topics from habitats to bay pollutants to nutrients and toxic chemicals) and view the full text of supporting material. An *enhanced* system needs additional sophistication including the integration of environmental data with secondary source information to help real problems associated with real tasks that may be executed by a range of people from citizens to policy makers and scientists. The emphasis is on examining patterns from the data, being receptive to the unexpected finding, while being focused on the contexts of use.

The Prototype

The prototype for the Bay of Fundy "collaboratory" is a starting point to sophisticated information access. At present, it contains the following elements:

1) The Greenstone software;
2) The BoFEP content.
3) The indexes to the BoFEP contents were used to enrich access;
4) A customized interface.

Open Source Software

The prototype was built using the award-winning open source digital library software, Greenstone, developed at the University of Waikto in New Zealand. Greenstone was designed for full text access, and has been used to date by many NGOs (see http://www.sadl.uleth.ca/nz/cgi-bin/library for samples) to provide access to similar types of materials. See http://www.greenstone.org for a complete description of the software. This software imports and indexes the full-text of documents in a variety of format such as *.doc*, *.pdf*, and *.htm*.

Content

The prototype contains the full text of the *BoFEP Proceedings* from 1996-2004, *Fundy Issues* to March 2006 and the *Coastal Forum Report of 2005*.

Each proceeding was a single PDF with a table of contents and an enriched back of the book index. Because we did not want the search for a single topic to retrieve an entire proceeding, all proceedings were sub-divided into individual units following the format used by the Association of Computing Machinery for proceedings that are contained in its Digital Library. Thus, each article served as a single item in this collection, and the entire set of items could be re-collated into its original proceeding. At the same time, the copyright and source information needed to be preserved, and thus a simple statement was posted to the bottom (or top) of each individual article indicating the Proceeding title to which it belonged.

While displaying a PDF is becoming the standard method of presenting a document on the Web, we also wanted to enable both a PDF download and Web presentation. Greenstone contains a PDF to HTML converter, but the conversion does not always result in perfectly formatted documents. For many of the proceedings, a MS Word (or

equivalent) version was also available, and that version was used to create an appropriately structured web document. In some cases only a PDF version was available. For one proceeding, only the print version was available. In this case, each paper was scanned as an image.

This preparation phase was a time-consuming process and we recommend that the digital version (e.g., .doc) of each paper be preserved in future for easier conversion to the Web.

Metadata – Physical Description
For each paper, a separate metadata record was created and an example is contained in Appendix B. These records contained the usual bibliographic information such as author, title, source and date which would later enable more flexibility in searching for each item. In addition, we included other fields such as Readership, e.g., for a general audience, to enhance future use.

Metadata – Conceptual Descriptors
Because the proceedings contained rich indexing and a Cumulative Index had been created by Rolston and Wells (2006), we de-constructed that index to assign each document a set of conceptual descriptors based on eight significant areas:

- Governance and management
- Habitats and ecosystems
- Fisheries and aquaculture
- Contaminants and pathogens
- Climate and climate change
- Energy and alternative energy
- Animals and Plants
- Place names

Each of these broad categories was further subdivided into a set of sub-categories that represent divisions of the concept. In addition, for each category, a set of general terms that are generic within the category (rather than specific to a sub-category) also emerged from this analysis. The conceptual classification scheme represented a substantial portion of the project.

Each paper belonged to at least one or more of these sub-categories or general terms. It was this low level structure that was used to create the category browser.

Information Access Interface
The interface to this collection contains a range of capabilities including a typical search engine and the ability to browse by a multi-layered set of environmental concepts that are directly related to the Bay:

1) The system provides a 'google' like search system.
The full contents of each document are searchable. In addition, the following metadata may be used to restrict or focus a search:

- author
- subjects
- year
- publisher
- type.

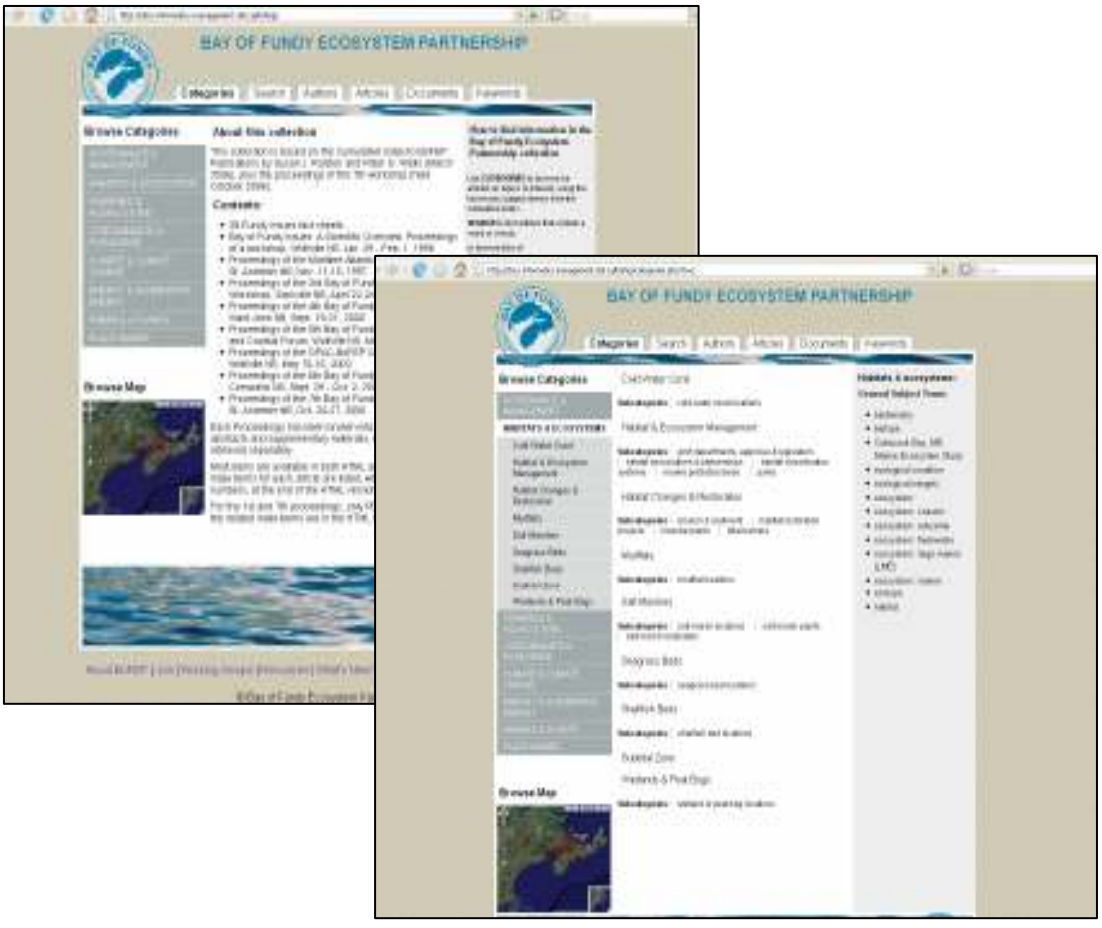
The full search capability of the Greenstone software is more fully described on its website.

Once a document is selected, the user can view (for most) the document in both Web or PDF form, as the illustration below shows. Once a search query is entered, a set of document titles are displayed. From this set a single document can be selected. The example shows a Web based presentation, but the document could be also presented as a PDF.

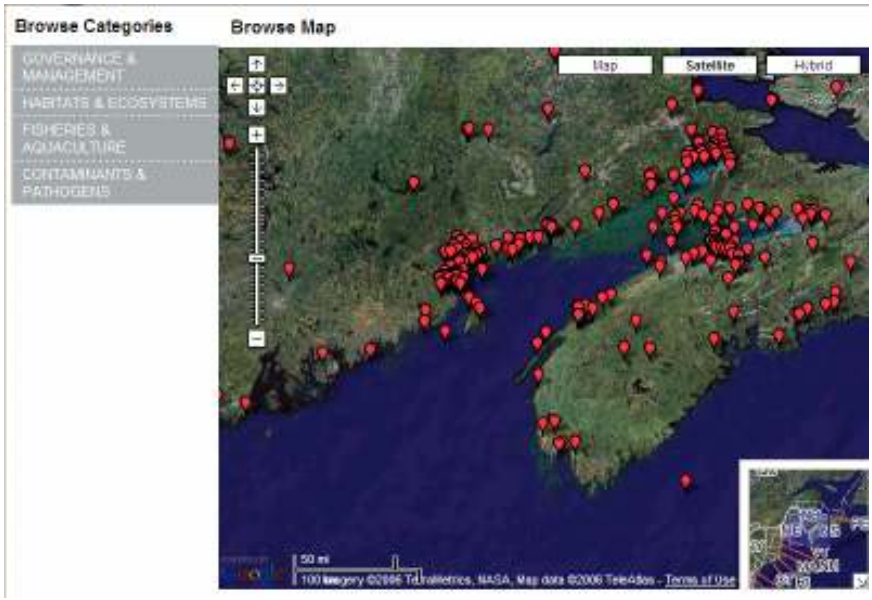
2) The enriched browser enables multiple pathways through the collection.
From the back-of-the-book indices, we created a special category browser that sits on top of the Greenstone software.

Prior to doing the development, we did a small user needs analysis using the members of the BoFEP Informatics Group members. The script that we used is contained in Appendix A. Working through these questions helped us to understand how this group thinks about and uses its information, and additionally how they process and access it. This analysis aided the process we used in categorizing the documents, particularly at the highest conceptual levels.

Initially, we identified four key categories that are important in the management of the Bay: Governance and Management, Habitats and Ecosystems, Fisheries and Aquaculture, and Contaminants and Pathogens. This list was deemed important in particular to policy makers. This set of four was later augmented based on a combination of responses from the group, and the types of documents that appeared in the BoFEP collection. The broad categories are illustrated to the left on the first illustration. The second shows the detail once the Habitats and Ecosystems section has been expanded. Once an item is selected then the same process as with the search described above is invoked. The selection leads to a list of titles that in sequence leads to a display of the article.



3) Spatial location is intrinsic to the problem domain. While an embedded geographic information system would be ideal, we simulated a modified effect by integrating the Google map tool for a demonstration of that capability. This resulted in a limited, but browseable map of the BOFeP resources around the Bay of Fundy and Gulf of Maine. The first image below shows all of the places around the Bay with related documents. By mousing over one of those red dots (as illustrated in the second image, the geographic location name, e.g., Victoria Beach, NS is displayed with a link that results in a search for that location. This is included to show what the potential for spatial access might be. See future work for more discussion on this point.



Current Status

The prototype as described is currently accessible from the Web at Centre for Management Informatics (http://docs.informatics.management.dal.ca/bofep/) and from the BoFEP (www.bofep.org) web sites. We have removed the integration with Google maps which we used in a demonstration to show the potential.

The prototype at the moment is in its infancy. It contains the document sets described earlier, as well as the search and browse capabilities. This represents a proof of concept.

Future Work

Our long term goal is to build an information and knowledge repository for the Bay of Fundy. It will be a decision making tool for use by communities of citizens, policy makers and researchers so that Fundy information is accessible and useful, indeed critical, for day to day decisions.

For the Bay of Fundy, we envisage being able to connect user communities to the data sources, including both the primary (the raw data) and secondary data (e.g. bibliographic, full text). The challenge which is still a significant research problem is in how to support specific tasks or jobs with the most pertinent data or information. , While we segmented the documents according to key priorities for policy makes, the output is still documents. Ideally, this should be key 'information chunks' that together aid in decision making/problem solving.

At present, we have including only textual documents – secondary evidence – which is essential to some policy makers and community members. But the expert in a topic also needs access to raw data such as that collected by buoys, weather systems, etc. An enhanced system would integrate that data with the secondary information. Integration wit existing systems such as GoMOOS (Gulf of Maine Ocean Observing System (see www.gomoos.org) would permit access to specific data on variables of interest (e.g. chlorophyll, at specific locations where automated instrument buoys have been collecting data for many years). This will facilitate comparisons between real time measurements and values reported in the research literature. Integrating databases with information retrieval systems also remains a challenging, but not insurmountable problem.

Ideally, the system will have in addition to the usual textual information and data, a means for visually accessing and presenting information by location. Geographic and the dynamic relationships between land and water and within water bodies mean that spatial context is an important factor in representing and interpreting information.

Finally, social networking on the Web has had a significant effect on everyday life. Those social connections are equally valid for geographically distributed communities such as BOFeP. Including other types of information such as experts on a topic and communication media such as presentations and documentaries would enrich the community of resources.

Acknowledgements

This project was supported with funding from the Bay of Fundy Ecosystem Partnership, Gulf of Maine Council on the Marine Environment, Environment Canada, and the Centre for Management Informatics at Dalhousie University. It is a project associated with the Fundy Informatics Working Group of BoFEP, and the authors' gratefully acknowledge the support and contributions of Pat Hinch and John Percy.

The Greenstone software was implemented by Joyce Goa; Tayze Mackenzie designed the interface and presentation of the category browser, and integrated it with the Greenstone system. The analysis of the keywords into a classificatory system was done by Ruth Cordes, while Susan Rolston with assistance from Donald Devoe prepared the content.

Reference

Rolston, S.J. and P.G. Wells. 2006. Cumulative Index to the Bay of Fundy Publications of BoFEP.

Note

A preliminary version of this report was presented at the Bay of Fundy Environmental Program Workshop in Fall 2006.

Appendix A. BoFIG User Need Analysis

Think back to the last time you needed to get material about the Bay of Fundy?

1) What was the problem that you were working on?

2) How would you characterize your problem?
_____ 1 learn about…
_____ 2 find out how to do something...
_____ 3 get advice..
_____ 4 look up facts..
_____ 5 find a solution…
_____ 6 find a tool/technique/instrument/method..
_____ 7 other:_____
_____ 8 other:_____

3) Describe the materials/information that you were hoping to find?

4) How would you characterize that material/information? By:

Characteristics of the Document/Information
_____ 1 Document type, e.g., journal article, statistics
_____ 2 Author
_____ 3 Date
_____ 4 Presentation, e.g., text or image

Geography, i.e., Elements of the Bay
_____ 5 Geographical location (named)
_____ 6 Geographic/geological feature
_____ 7 Species/genus
_____ 8 'Chemical' element

Organizational involvement
_____ 9 Specific research projects (e.g., an ACAP project)
_____ 10 Organizations, e.g., universities, research labs, not-for-profit, government agencies
_____ 11 Management/policy initiatives (e.g., Oceans Action Plan, UNESCO biosphere initiative, IMO)

Types of Research
_____ 12 Broad category of environmental research (e.g., birds, salt marshes, tidal flow)
_____ 13 Socio-economic issues (e.g., community participation)
_____ 14 General marine-related issues (e.g., shipping, tourism)

Process
_____ 15 Technique, i.e., how things are done (e.g., best management practices, bathymetry, MEQ)

Appendix B. Metadata Record

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE DirectoryMetadata SYSTEM
"http://greenstone.org/dtd/DirectoryMetadata/1.0/DirectoryMetadata.dtd"
>
<DirectoryMetadata>
...
<FileSet>
<FileName>BOFEP7-2006-037.htm</FileName>
   <Description>
     <Metadata name="articleID">51736</Metadata>
     <Metadata name="DocType">article</Metadata>
     <Metadata name="Readership">general</Metadata>
     <Metadata name="Series">Bay of Fundy Ecosystem Partnership
Technical Report No. 3</Metadata>
     <Metadata name="Title">Enhancing Information and Knowledge of the
Bay of Fundy</Metadata>
     <Metadata name="Page">37</Metadata>
     <Metadata name="Year">2007</Metadata>
     <Metadata name="ISBN">978-0-9783120-0-8</Metadata>
     <Metadata name="Publisher">Bay of Fundy Ecosystem
Partnership</Metadata>
     <Metadata name="Proceedings-Series">Proceedings of 7th workshop,
St. Andrews NB, Oct. 24-27, 2006</Metadata>
     <Metadata name="Proceedings">Challenges in Environmental
Management in the Bay of Fundy-Gulf of Maine. Proceedings of the 7th
Bay of Fundy Science Workshop, St. Andrews, New Brunswick, 24-27
October 2006</Metadata>
     <Metadata name="Language">English</Metadata>
     <Metadata name="PDF">/gsdl/collect/bofep/pdf/WG/BOFEP7-2006-
037.pdf</Metadata>
     <Metadata name="Session">Paper Presentations: Session 2:
Environmental Issues</Metadata>
     <Metadata name="end page">39</Metadata>
     <Metadata name="Author" mode="accumulate">Toms, Elaine
G.</Metadata>
     <Metadata name="Author" mode="accumulate">Cordes, Ruth
E.</Metadata>
     <Metadata name="Author" mode="accumulate">Gao, Joyce</Metadata>
     <Metadata name="Author" mode="accumulate">MacKenzie,
Tayze</Metadata>
     <Metadata name="Author" mode="accumulate">Rolston, Susan
J.</Metadata>
     <Metadata name="Author" mode="accumulate">Hinch, Patricia
R.</Metadata>
     <Metadata name="Author" mode="accumulate">MacDonald, Bertrum
H.</Metadata>
     <Metadata name="Author" mode="accumulate">Wells, Peter
G.</Metadata>
     <Metadata name="Authors">Toms, Elaine G.; Cordes, Ruth E.; Gao,
Joyce; MacKenzie, Tayze; Rolston, Susan J.; Hinch, Patricia R.;
MacDonald, Bertrum H.; Wells, Peter G.</Metadata>
     <Metadata name="Subject" mode="accumulate">BoFEP (Bay of Fundy
Ecosystem Partnership): Fundy Informatics Working Group</Metadata>
```

```
     <Metadata name="Subject" mode="accumulate">Chesapeake
Bay</Metadata>
     <Metadata name="Subject" mode="accumulate">geographic information
system (GIS)</Metadata>
     <Metadata name="Subject" mode="accumulate">governance</Metadata>
     <Metadata name="Subject" mode="accumulate">Gulf of Maine Council
on the Marine Environment (GOMCME)</Metadata>
     <Metadata name="Subject" mode="accumulate">Gulf of Maine Ocean
Observing System (GoMOOS)</Metadata>
     <Metadata name="Subject" mode="accumulate">information and
data: access</Metadata>
     <Metadata name="Subject" mode="accumulate">information and
data: technology</Metadata>
   </Description>
</FileSet>
...
</DirectoryMetadata>
```