



WHITE PAPER

Capture of Exploration Enterprise Web Sites

Developed at the
NASA Goddard Space Flight Center

by

Nikkia Anderson, Gail Hodge, (Information International Associates) &
Ed Rogers, Systems Management Office, Code 300

May 17, 2004



Introduction

An increasing amount of scientific and technical information is created digitally and disseminated via the Web. NASA has millions of internal and external web pages. Web pages, particularly those with a public presence, are highlighted in the efforts of the Interagency Committee on Government Information which is responding to the requirements of the E-Government Act of 2002. Web sites are growing in number and in complexity. In addition to traditional HTML text, a site may include video and audio clips, dynamic content based on a database query, digital still images, animations, 3D models, and PDF documents. It can be anticipated that much of NASA's internal and external information will be shared via the Web and in a variety of formats.

In response to this changing environment, NASA Goddard Space Flight Center's Library and the GSFC Knowledge Management Office with support from the GSFC Director's Discretionary Fund has developed a system for capturing web sites of significant scientific and technical interest in support of knowledge management initiatives.

The goal of the Web Capture Project (Figure 1) is to provide a web application that captures web sites of long-term scientific and technical interest, stores them, extracts metadata, if possible, and indexes the metadata in a way that the user can search for relevant information. During the past year, the GSFC Library and Information Services Branch captured over 230 Web sites (pages that have intellectual cohesiveness) with more than 91,000 pages.

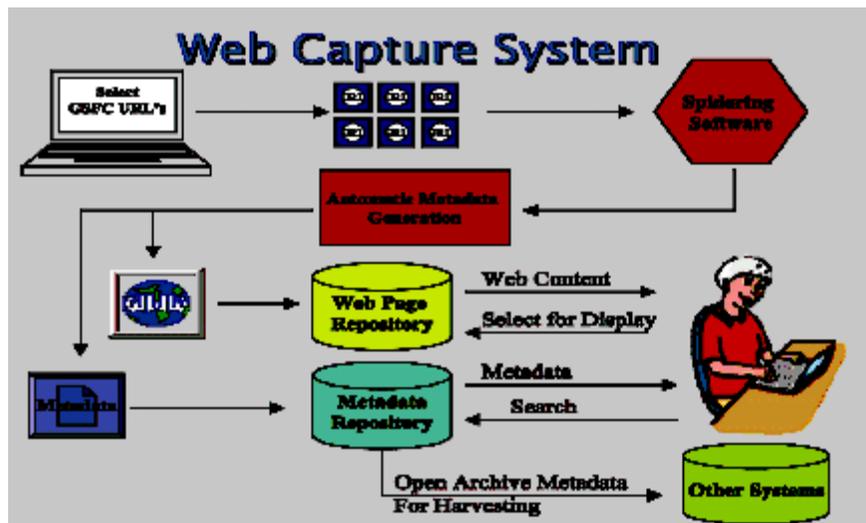


Figure 1: Web Capture System Flow

Web Site Selection

enhancing of the Goddard Core Elements (Figure 2). Controlled vocabulary terms, a text description and free text keywords are added manually during the review process. The competency is linked to two controlled vocabularies, the NASA-wide Taxonomy [Dutra & Busch, see: <http://nasataxonomy.jpl.nasa.gov/>] and the Earth Observing System Taxonomy developed internally at GSFC. Each taxonomy has a controlled vocabulary that appears on the form as a drop-down menu.

Searching the Metadata Records

The system uses an open source search engine, Lucene, to index and search the metadata. In this context Lucene indexes and searches only the metadata from database but Lucene could be set up to index the full text of the pages (HTML, XML, etc) or other document formats (Word, PDF, etc.) if tools to extract text from them exist.

The search form allows the user to enter terms and select specific metadata elements of the Goddard Core on which to search (Figure 3). Current searchable fields include: title, description, keyword, competency, creator (author) and organization code.



Figure 3: Search Page

When a search is executed, the results are paged and displayed in a table that contains basic information to allow the user to evaluate the resulting hits (Figure 4). In the prototype, these fields include Title, Creator, Competency (Subject Discipline), and Organization Code.

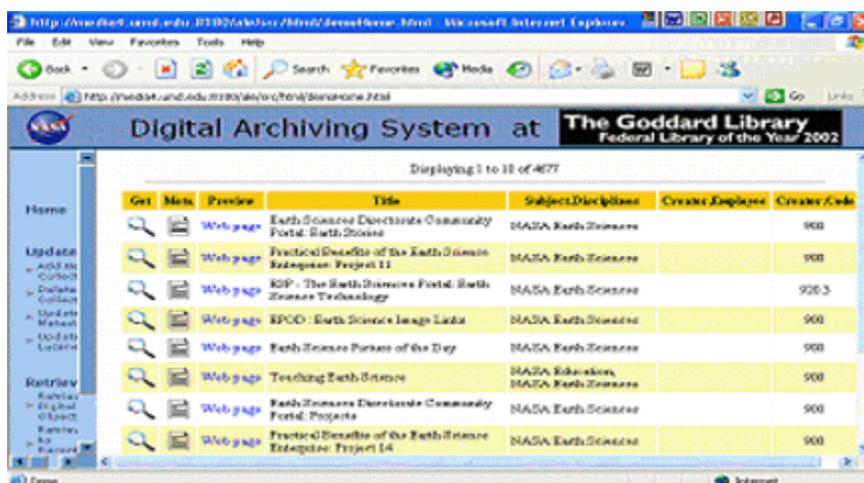


Figure 4: Results Page

The user can display the web site by clicking on the magnifying glass and opening the page in a new resizable window that allows full navigation of the portion of the site that is in the GSFC domain. The full metadata record can be displayed or a small thumb-nail size of the Web page can be previewed.

Outstanding Challenges and Opportunities

Through the work performed to-date, the GSFC Library and Information Services Branch has developed a research agenda required to improve the capturing of web sites. These include more automatic creation of metadata, capturing of the dynamic and deep web, and version control. The GSFC Library and Information Services Branch continues to search for innovative ways to improve the capture process, including monitoring the work of national and international projects in this area.

Applicability to NASA Program and Project Management

In the new digital environment NASA has the unique opportunity to create systems that manage information throughout the program/project life cycle and in a variety of formats. As the timing between the activity and the next round of technological development and innovative grows ever shorter, there will be an increased need to capture, preserve and make available web sites and other web-based materials to support a learning organization. NASA should consider an Agency web capture initiative to preserve this valuable source of information for the public as well as future internal needs.