Date : 26/05/2008

Archiving Foreign Government Statistical Web Sites for all at Indiana University Libraries

Andrea Singer
Indiana University Bloomington Libraries
Bloomington, IN., USA

Meeting: **130. Government Information and Official Publications**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/iv/ifla74/index.htm

Abstract:

Context: In his overview of web archiving initiatives at Global Resources Network (http://www.crl.edu/grn/index.asp) institutions, James Simon of the Center for Research Libraries lists ten active initiatives for capturing and curating web sites, including Archive-It's subscription-based archiving service. Indiana University Libraries is among the institutions that began subscribing to Archive-It (http://www.archive-it.org/) in 2006 in order to capture, preserve, and search web sites, without needing technical support or infrastructure.

This paper will focus on a government information web site archiving project – one of three collections being archived at Indiana University. Two of the collections are obviously clearly related to local preservation needs: the web sites of Indiana University, and selected state of Indiana and local government web sites. The third collection, which is the subject of the paper, consists of web sites of national level governmental statistical agencies outside the European Union, Australia, Canada, and the USA.

For more than half a century at Indiana, selection and acquisition of documents produced at the national level by governments outside the United States has emphasized key genres for social science and historical research: statistical abstracts, legislative reports, development plans, and education reports, for example. As a greater proportion of this material became available only on the web, we chose to devote some of our budget to preserving web sites rather than specific documents through Archive-It.

Cooperation: The archived web sites are freely available and searchable by key word or URL at the Archive-It web site. Institutions can add supplementary search interfaces, web sites, and user aids outside the Archive-It interface, and determine how often to capture, disable, or add new sites.

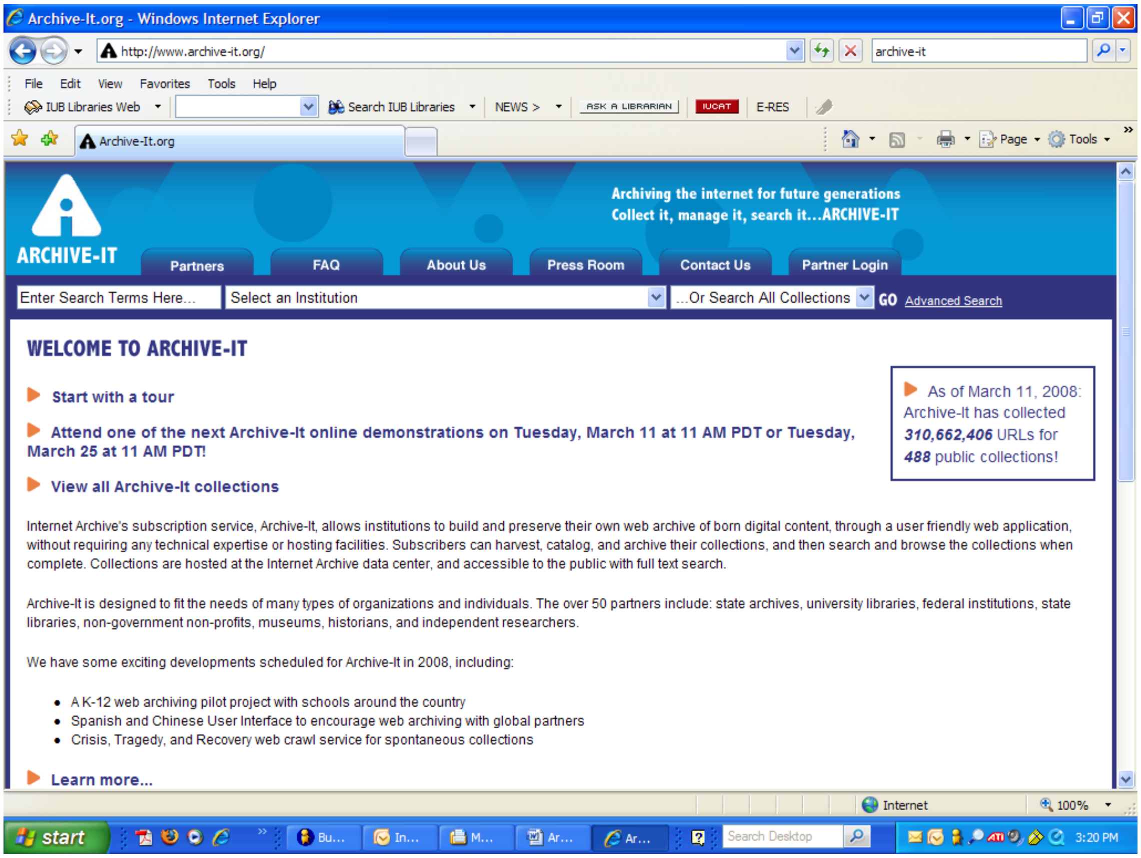
Like much at Indiana, work on developing this project took place in a committee or working group. This one had participation from Cataloging, the University Archives, and Collection Development staff. Technology staff was not involved, and a subject librarian maintains accompanying web resources and seeds for each collection.

Supplementary finding aids include records for our collections at the top level in our local catalog and OCLC WorldCat, paper collection brochures, and web links such as these:
http://www.libraries.iub.edu/index.php?pageId=4302
http://www.archive-it.org/collections/317
http://www.libraries.iub.edu/index.php?pageId=4981

Wider issues of current cooperative concern revolve around making certain we are not duplicating the efforts of others in this area. Selection of seeds which are not being archived elsewhere, seeking information about registry plans for URLs, and other aspects of cooperation will be discussed.

Biographical Information:
Andrea Singer is the Foreign Documents Librarian and Bibliographer for India and Tibetan Studies at Indiana University Libraries, Bloomington, Indiana. She is the curator of the Foreign Government Statistical Web Sites Collection.

Thank you for the opportunity to share this report of an experiment in web-archiving. As I began writing this paper in March I realized that some of the web sources I had consulted while preparing the abstract earlier in 2008 had already moved or disappeared. That evidence caused me to minimize inclusion of specifics such as search techniques, so that on presentation to you in August, we would all be in synchronization with the current web as much as possible. The header at the top of the Archive-It welcome page in March, 2008 read: ”Archiving the Internet for future generations Collect it, manage it, search it…ARCHIVE-IT”, and presented a key word search box for an institution or all collections.



WELCOME TO ARCHIVE-IT

- Start with a tour
- Attend one of the next Archive-It online demonstrations on Tuesday, March 11 at 11 AM PDT or Tuesday, March 25 at 11 AM PDT!
- View all Archive-It collections

As of March 11, 2008: Archive-It has collected *310,662,406* URLs for *488* public collections!

Internet Archive's subscription service, Archive-It, allows institutions to build and preserve their own web archive of born digital content, through a user friendly web application, without requiring any technical expertise or hosting facilities. Subscribers can harvest, catalog, and archive their collections, and then search and browse the collections when complete. Collections are hosted at the Internet Archive data center, and accessible to the public with full text search.

Archive-It is designed to fit the needs of many types of organizations and individuals. The over 50 partners include: state archives, university libraries, federal institutions, state libraries, non-government non-profits, museums, historians, and independent researchers.

We have some exciting developments scheduled for Archive-It in 2008, including:

- A K-12 web archiving pilot project with schools around the country
- Spanish and Chinese User Interface to encourage web archiving with global partners
- Crisis, Tragedy, and Recovery web crawl service for spontaneous collections

Learn more...

The Context:

The Internet Archive offers a subscription-based system, Archive-It, for archiving web sites by using specific URLs as keys, or, in Archive-it terminology, seeds. Archive-It partners, participants, or subscribers select web sites which are grouped in collections, and control the frequency of the harvest of selected web sites. They also have opportunities to limit the searches of a particular URL in specific ways.[1] The collections are harvested, housed, and searched by Archive-It, and no technical expertise is required of partners, who are free to focus on content.

Indiana University Libraries - Bloomington, the research library of a large public university in the mid western United States, began subscribing to the Archive-It service in 2006 after an initial planning period beginning in 2005. We have selected URLs to build three collections. In the time-honored tradition of "thinking globally, acting locally" two are collections of great local interest. The "Indiana University Web Sites" collection, managed by staff at the University Archives, is the collection we feel the most responsibility to preserve. "Indiana: State and Local Governments", managed by the librarian for state and local government

[1] The URLs in the National Government Statistical Web Sites collection have not been purposely limited. If a capture is not blocked or limited by the web site selected for archiving, the default capture extends two places.

information, is also important to the history of our state and region. (We believe there is no overlap with other archiving programs except the general Wayback Machine[2] for these two collections .) The third collection is a selective group of national government web sites authored by statistical agencies worldwide. The number of sites chosen for archiving has varied from seventy to approximately two hundred over the short time of the experiment. The focus on sites outside Western Europe, Australia, New Zealand, Japan, and North America—countries or regions which have fledgling or well-developed web archiving experiments underway[3] -- has remained constant.

The choice of sites has grown from collection development philosophies and policies which have been in place at Indiana for many years for collecting foreign-to-the-United States government documents and information. The notion that social scientists, historians, and other researchers rely perennially on certain types of information collected by governments such as statistical publications, censuses, development plans, and reports on education, the environment, and other topics of sustained research and interest has been the basis of our selection of print resources, since covering the whole range of government publishing for many countries is not possible.[4] (The best explication of these generic publication types and their print locations is probably Gloria Westfall's *Guide to Official Publications of Foreign Countries*[5], which she characterizes as "the result of international cooperation and joint efforts of many librarians from around the world.")

Over time, as governmental web sites began to supply the most up-to-date information in increasingly fuller renditions, we joined the librarians who routinely used web reference sites to supplement print resources. A powerful freely available reference tool which we eagerly added to our toolbox was Gunner Anzinger's *Governments on the WWW*[6]*.* The author maintained this site from 1995 until mid-2002. In addition, we asked catalogers to add records for selected governmental web sites to our local catalog, IUCAT, and to OCLC WorldCat. The records included PURLs, but, of course, could not link to information which had been removed from the web. (Selection and cataloging in this selective manner is clearly expensive, because of the attention required by highly trained staff. Hence we were eager to experiment with other methods of capturing and preserving digital content.)

The abstract of this paper includes URLs for several web pages associated with the Archive-It collection we are developing, including an overview, a searching aid, and the Archive-It search page for the collection. An IU Libraries page which reveals more about the context of the collection in our home institution couches the links to the archival collection in a not-

[2] The Wayback Machine has archived web sites since 1996. See http://www.archive.org/web/web.php (accessed April 1, 2008).

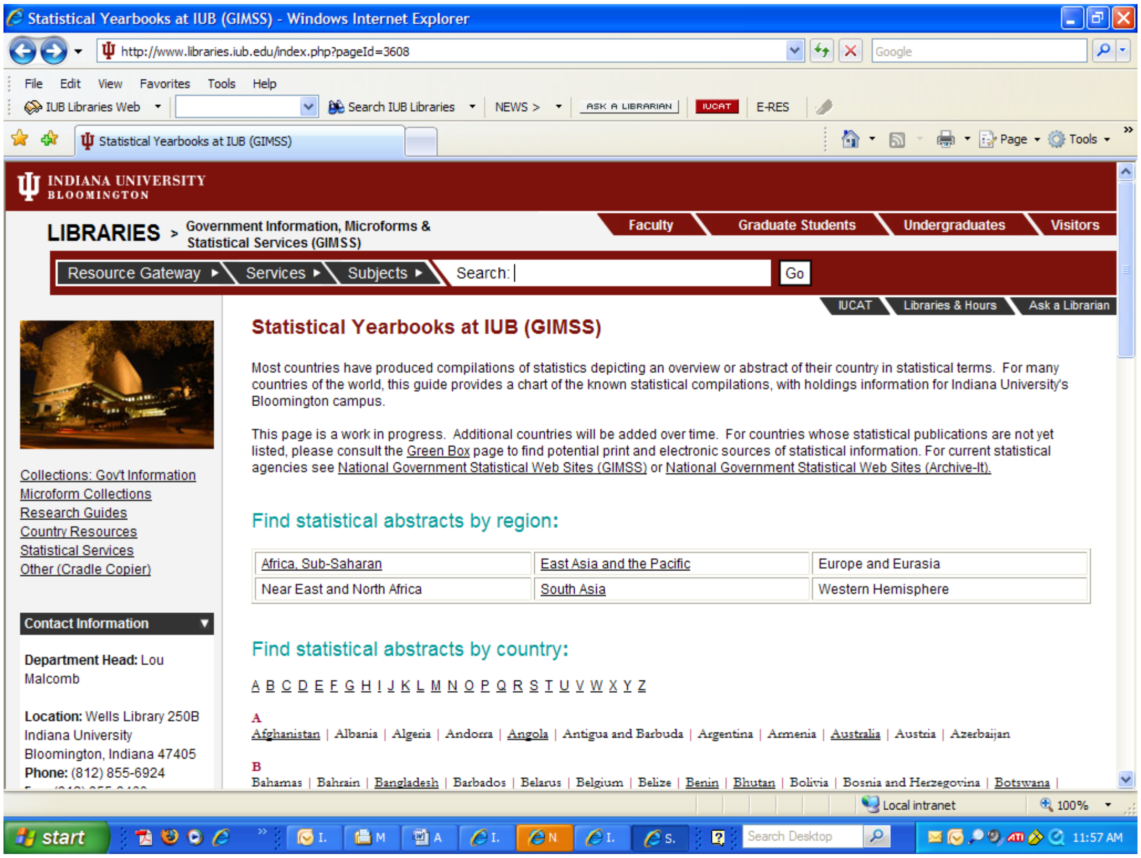
[3] The initial list of possible sites was obtained at the U.S. Census Bureau's list of Statistical Agencies (International) in 2006. See http://www.census.gov/main/www/stat_int.html (accessed April 1, 2008).

[4] Indiana University has strong area studies programs, so basic sources are supplemented in depth for some parts of the world.

[5] Gloria Westfall, ed. *Guide to Official Publications of Foreign Countries.* (Bethesda, Md.: CIS, 1990.) American Library Association/Government Documents Round Table. Second ed., v.

[6] Still available in English and German, this source provides quick links to governmental web sites by region, country, and selected category of information such as "Institutions in the area of Statistics", or "Parliaments", for example. See http://www.gksoft.com/govt/ (accessed April 1, 2008).

yet-completed finding aid for our statistical yearbook collections.[7] Links to our local finding aids and the Archive-It collection page are included.



Statistical Yearbooks at IUB (GIMSS)

Most countries have produced compilations of statistics depicting an overview or abstract of their country in statistical terms. For many countries of the world, this guide provides a chart of the known statistical compilations, with holdings information for Indiana University's Bloomington campus.

This page is a work in progress. Additional countries will be added over time. For countries whose statistical publications are not yet listed, please consult the Green Box page to find potential print and electronic sources of statistical information. For current statistical agencies see National Government Statistical Web Sites (GIMSS) or National Government Statistical Web Sites (Archive-It).

Find statistical abstracts by region:

Africa, Sub-Saharan	East Asia and the Pacific	Europe and Eurasia
Near East and North Africa	South Asia	Western Hemisphere

Find statistical abstracts by country:

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A
Afghanistan | Albania | Algeria | Andorra | Angola | Antigua and Barbuda | Argentina | Armenia | Australia | Austria | Azerbaijan

B
Bahamas | Bahrain | Bangladesh | Barbados | Belarus | Belgium | Belize | Benin | Bhutan | Bolivia | Bosnia and Herzegovina | Botswana |

Department Head: Lou Malcomb

Location: Wells Library 250B
Indiana University
Bloomington, Indiana 47405
Phone: (812) 855-6924

Cooperation:

Internal Project:

This description of cooperation begins with the initial decision by Indiana University Libraries-Bloomington administrators to join other research libraries in an RLG project with Archive-It in 2005.[8] Internally the project was headed by the Libraries' Director of Collection Development, and a working group from Archives, Cataloging, Collections, and Public Services units. Support was and still is provided by the assistant to the Director of Collection Development. (The composition of the group, which did not include staff from the Digital Library Program or Library Information Technology units is significant because it confirms that expertise in technical areas is supplied by Archive-It.) By 2006, the planning group had decided how to organize and represent Archive-It collections, and was ready to begin subscribing for the three collections already described.

[7] Indiana University Libraries – Bloomington."Statistical Yearbooks at IUB (GIMSS)" See http://www.libraries.iub.edu/index.php?pageId=3608 (accessed March 13, 2008).

[8] Originally founded in 1974 by The New York Public Library and Columbia, Harvard, and Yale universities, RLG Programs became part of OCLC in July 2006. See http://www.oclc.org/programs/about/default.htm (accessed April 1, 2008).

The Working Group decided to describe all three collections at the top level – by the name of each specific collection -- in IUCAT and OCLC WorldCat. Metadata creation was minimal, and internal users were alerted to the presence of the collections through standard descriptive web pages as well as a page devoted to search tips for each individual collection. As is the case for other Indiana University Bloomington library collections, we also created and continue to distribute a paper brochure to describe the collections. The Libraries also featured the collections on its home page, and in the 2006/07 annual report, which was distributed to Indiana University faculty, administrators, and library supporters.

At the same time, limited assessment of the collections was carried out by the individual curators. Links were added or disabled based on information from the initial captures or crawls, and we experimented with crawl frequency, consulting periodically to ensure that we were not overspending our budget for this project.[9]

Other instances of cooperation at the institutional level included participation in the first Archive-It partners meeting in San Francisco in October, 2007, and current cooperation for this particular collection in a broader experiment in open searching through an experimental partnership between Archive-It and Oregon State University's *LibraryFind* project.[10]

A Wider Issue:

Many issues concern all participants in web archiving projects. In addition to those which relate to selection and technology, prominent concerns include preservation over the long term, avoiding duplication of effort, and quality control and appraisal.

Small scale-- perhaps even micro scale-- contributions to worldwide archiving efforts, including the National Government Statistical Web Sites collection, archive tiny numbers of URLs in contrast to the scope of the projects on which national libraries and others have embarked. Curators of these small collections often have many other non-digital responsibilities and are certainly not experts in web archiving. (I expect most of us glean information about projects, techniques, and methods for appraisal from our host organizations, or from the open web, which is a constantly changing cooperative effort.)

Many tools are available on the open web including Harvard University's Office for Information System's page on "Web Archiving Resources".[11] Here sections on lists, IP rights, metadata and cataloging, web archives, workshops and conferences, and general web archiving information are collected. Other work on pilot efforts to describe requirements for registries of digital information is accessible as well on the open web.

[9] The decision to limit captures to annual crawls, for example evolved because of internal economic factors. Currently funded centrally, the desire to sustain the project if necessary with government documents collection funds has influenced decisions about content.

[10] Archive-It invited institutions to participate in this experiment on a collection by collection basis in February, 2008.

[11][Harvard University] Office for Information Systems. "Web Archiving Resources". http://hul.harvard.edu/ois/projects/webarchive/resources.html (accessed March 13, 2008).When accessed it had last been modified on October 24, 2006.

Yet from the point of view of any user who might be trying to work backward from current information to locate earlier web sites, one of the challenges is locating information on whether a particular web site has been archived anywhere. The Wayback Machine, which I mentioned early in this presentation, supplies that information for the sites it archives.

The following brief discussion, which concludes this presentation, muses on the value of cooperatively creating a registry of archived websites, which could benefit archivists as well as users.

In the case of the National Government Statistical Web Sites collection, core users are those who need help with hard-to-locate statistical information in our local institution. We use the collection as we would any other tool in the toolbox, and we expect external users will discover available information on the Archive-It web site in a search for archived material, rather than by needing to know the particulars of access at Indiana. However we would be happy to contribute to an open registry, which might require only a URL for retrieval, the URL of the archived site, and the date(s) of web capture(s) to be useful for searchers and developers.

I hope audience members will share their ideas on this topic during the question period, or after the session.

Thank you.