OHIO ELECTRONIC RECORDS COMMITTEE
http://www.ohiohistory.org/ohiojunction/erc/

GUIDELINES FOR MANAGING WEB SITE CONTENT

SECTION 1 - SCOPE:

These guidelines are appropriate for State of Ohio agencies, boards and commissions. Other governmental entities may also wish to follow these guidelines as appropriate.

SECTION 2 - INTENT AND PURPOSE:

The intent of these guidelines is to raise awareness of and provide and explain the currently available requirements, guidelines and best practices for managing and preserving web resources that meet the criteria for records as defined by the Ohio Revised Code.

These guidelines have a two-fold purpose. First, they are intended to assist state agency employees in complying their management of web resources with Ohio public records law.

Second, the guidelines promote best practices and suggestions that facilitate the effective creation, management, and retention of web resources as public records.

Because technology is changing so rapidly, records management tools are not keeping pace with the tools we use to create and distribute information. Consequently, these guidelines will offer the best advice that is currently available. These guidelines will be updated as more research is performed and new tools are developed to enable records management and preservation in the electronic environment.

SECTION 3: INTRODUCTION

> ➤ **Agency web sites contain and generate records.**

Both internal and public web sites contain records, including reports, guidelines and advice, and policies and procedures. Often these records already exist in other formats. Nevertheless, records are increasingly being created for direct placement on web sites. This content may take the form of a static document, or may be created dynamically from a database behind the web site. Unless specific policies and procedures are in place, many web-based records documenting the business of the agency will not be scheduled and retained appropriately.

Web sites can also be used as the technology to underpin interactions between an agency and its clients. Agencies must ensure that records documenting these transactions are scheduled and retained appropriately.

> ➤ **Web resources as publications**

The dual purpose of a state web site is to:

> Publish information

> Facilitate the provision of products and services

Due to the inherently "public" nature of web sites, many of the discrete information resources agencies include on the web are considered to be publications. As is the case in the paper-based environment, publications are also records and are subject to specific retention and distribution requirements.

> ➤ **Maintaining Web Resources: A Shared Responsibility**

Because of the differing mandates of the State Archives of Ohio and the State Library of Ohio, webmasters and content creators need to be aware of an important distinction in their web content. This distinction is between "records" and "publications"

> "Records" includes any document, device, or item, regardless of physical form or characteristic, created or received by or coming under the jurisdiction of any public office of the state or its

political subdivisions, which serves to document the organization, functions, policies, decisions, procedures, operations, or other activities of the office." (ORC 149.43)

Publications include, but are not limited to, reports (annual, technical, research, statistical), directories, pamphlets, brochures, fact sheets, laws, rules, handbooks, manuals, bulletins, circulars, forms, newsletters, press releases, maps, charts, multi-media files that are intended for public use and distribution by any department, division, bureau, board or commission of the state government, regardless of format. Publications exclude information that is for strictly internal administrative or operational purposes, having no public interest, educational, or historic value.

Publications are a subset of records. Like records, publications should be retained by the agency according to a retention schedule. However, publications are also maintained and preserved by the State Library while records are maintained and preserved by the State Archives of Ohio if they have enduring historical value. Records that do not have enduring historical value must be maintained by the originating agency for the duration of the appropriate retention periods.

SECTION 4 – DEFINITIONS:

Dublin Core Metadata: Metadata is, in its most simple definition, data about data. The Dublin Core Metadata Initiative (DCMI) is dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems.

Publication: Publications include, but are not limited to, reports (annual, technical, research, statistical), directories, pamphlets, brochures, fact sheets, laws, rules, handbooks, manuals, bulletins, circulars, forms, newsletters, press releases, maps, charts, multi-media files that are intended for public use and distribution by any department, division, bureau, board or commission of the state government, regardless of format. Publications exclude information that is for strictly internal administrative or operational purposes, having no public interest or educational or historic value.

Record: "Records" includes any document, device, or item, regardless of physical form or characteristic, created or received by or coming under the jurisdiction of any public office of the state or its political subdivisions, which serves to document the organization, functions, policies, decisions, procedures, operations, or other activities of the office." (ORC 149.43)

Recorded information, in any form, including data in computer systems, created or received and maintained by an organization or person in the transactions of business or the conduct of affairs and kept as evidence of such activity.

Record Series: Records that are created for a common purpose, logically relate to one another, are filed or maintained together, and that have the same retention period.

Retention Schedule: A document that establishes the length of time a particular record series must be retained by the originating office in its particular format(s).

Snapshots: Full and accurate record copies of an agency's public web resources captured at particular points in time.

Web Resource: A discrete file or group of files available via the World Wide Web that is created and managed as a unit to convey information about a particular subject.

Web site: A collection of electronic files, usually under common administrative control, linked together and made accessible to the public via the World Wide Web.

SECTION 5 - LEGAL REQUIREMENTS:

5.1 LEGAL REQUIREMENTS FOR ALL RECORDS

The Ohio Revised Code (ORC) includes no specific definition for web sites, however ORC Section 149.011(G) provides the following definition:

> "Records" includes any document, device, or item, regardless of physical form or characteristic, created or received by or coming under the jurisdiction of any public office of the state or its political subdivisions, which serves to document the organization, functions, policies, decisions, procedures, operations, or other activities of the office."

Clearly, a web site is a document or item created or received by a public office. Whether the web site documents the organization, functions, policies, decisions, procedures, operations or other activities is the deciding factor as to its status as a record.

Web sites that meet the criteria of the definition of a record must be scheduled on a retention schedule and retained for the appropriate time period before disposition. Scheduling web sites is discussed in detail below.

Web sites that meet the criteria of the definition of a record, as defined in the Ohio Revised Code, may also be classified as public records as defined in the Ohio Revised Code Section 149.43, which states in part:

> "Public record means any record that is kept by any public office, including, but not limited to, state, county, city, village, township, and school district units, except that public record does not mean any of the following: medical records…"

> All public records shall be promptly prepared and made available for inspection to any person at all reasonable times during regular business hours. Subject to division (B)(4) of this section, upon request, a public office or person responsible for public records shall make copies available at cost, within a reasonable period of time. In order to facilitate broader access to public records, public offices shall maintain public records in such a manner that they can be made available for inspection in accordance with this division.

As with any format, a web site is considered a public record unless it falls under one of the exceptions listed in Section 149.43. Public records must be maintained and made accessible to the public upon request through the appropriate retention period.

5.2 LEGAL REQUIREMENTS FOR PUBLICATIONS

The current legal requirements for publications do not adequately address web-based publications. Section 7 of these Guidelines is intended to provide advice to state agencies that will assist them in ensuring that their web-based publications are preserved. ORC 149.11 states:

"Any department, division, bureau, board, or commission of the state government public use and distribution, which publication is reproduced by duplicating processes such as mimeograph, multigraph, planograph, rotaprint, or multilith, or printed internally or through a contract awarded to any person, company, or the state printing division of the department of administrative services, shall cause to be delivered to the state library one hundred copies of such publication, subject to the provisions of section 125.42 of the Revised Code."

The state library board shall distribute the publications so received as follows:

(A) Retain two copies in the state library;
(B) Send two copies to the document division of the library of congress;
(C) Send one copy to the Ohio historical society and to each public or college library in the state designated by the state library board to be a depository for state publications. In designating which libraries shall be depositories, the board shall select those libraries which can best preserve such publications and which are so located geographically as will make the publications conveniently accessible to residents in all areas of the state;

SECTION 6 - RECORDS MANAGEMENT FOR WEB-BASED RECORDS:

6.1 BEST PRACTICES FOR SCHEDULING WEB-BASED RECORDS

SCHEDULING ACCORDING TO CONTENT

Web sites are not considered a record series or category. They are a means of publishing information. Like paper or microfilm, web sites are the medium by which this type of record is created and maintained. Just as an agency cannot schedule all paper or microfilm records together under a single retention period, an agency generally cannot simply schedule a web site as a record series. Rather, retention or disposition of web site files must be related to the information they contain or the purpose they serve. The content of the various pages on a web site may vary considerably, and therefore, this content must be evaluated to determine the length of time the web resource must be retained.

SCHEDULING ACCORDING TO FORMAT

Many web-based records are maintained in more than one format. In other words, a resource may be published on the web site and be maintained in a paper format by the agency. In these cases, the retention schedule should reflect the multiple formats and the agency might want to adopt differing retention periods based on format.

By using this method, agencies can commit to maintaining paper (or microfilm) copies of resources for longer periods of time than they commit to retaining electronic, web based copies of resources. Agencies are strongly cautioned, however, that printing out web pages is not a solution for all web resources. This option works best for distinct web resources that are routinely maintained in more than one format.

BEST PRACTICE 1

Government agencies must identify records that exist on their web site(s), or which are about to be placed online. They must ensure that these records are included on appropriate records retention schedules and that procedures to implement the retention periods are established.

BEST PRACTICE 2

Government agencies should ensure that full and accurate records of their web resources are captured and maintained as specified in the retention schedules. Agencies must capture full and accurate records of web-based transactions.

BEST PRACTICE 3

Government agencies should make and keep records that accurately document their public web resources over time, so that it is possible to reliably establish the content of their web sites for the duration of the appropriate retention period.

BEST PRACTICE 4

Government agencies are encouraged to develop their own web-based records management policy.

BEST PRACTICE 5

Government agencies should develop a program for conducting regular reviews to identify the effectiveness of their records management procedures for web-based resources. Conducting a review is especially important when the web sites is modified or when business processes change.

6.2 GUIDELINES FOR MANAGING WEB-BASED RECORDS

A single web site may be made up of a variety of resources with differing degrees of technical complexity. Applying the concepts of records scheduling, management, disposition, and preservation will increase in complexity in tandem with the increase in the technical complexity of the web site.

There is no generic solution for creating and maintaining records of web-based activity. The best option will depend on the outcome of an analysis of the particular circumstances. Each agency should assess a number of factors, including:

- The type of web-based activity—information dissemination or transaction or both;
- The complexity, diversity and sophistication of the web-based activity;
- The frequency and regularity of change to the web-based activity;
- The frequency of challenges to the validity of the information on the website;
- The level of risk and public visibility of the agency;
- The agency's record retention requirements;
- The agency's technological environment; and
- The availability of resources.

6.2.1 STATIC WEB SITES

In its most basic form, a web site may be nothing more than a collection of static documents sitting in folders on a server and tied together with hyperlinks. The only interactivity provided by static sites is in the links that enable movement from one document to another or from one part of the site to another.

SCHEDULING

Static sites are relatively simple to schedule. Steps included in the process include:

1. Identify distinct record series
2. Is the series included on an existing agency retention schedule or in the General Schedules?
3. If the series is included on an existing schedule, you should:
   a. Determine if the retention period if appropriate. Revise as needed
   b. Determine if the web-based format is included on the retention schedule. If it is not, revise the retention schedule to include web-based format.
4. If the series is not included on an existing schedule, you should:
   a. Determine if the series is maintained in other formats (paper or microfilm)
   b. Determine appropriate retention period
   c. Create and submit retention schedule

IMPLEMENTING THE RETENTION SCHEDULE REQUIREMENTS

1. Work with webmasters and content creators to create procedures for updating files and maintaining deleted files for the appropriate retention period. Include appropriate metadata in web pages that includes the date the page was made live; the date the page was removed from the web; and the proposed disposal date.
2. Move updated and deleted files to alternative storage medium for the length of the retention period.

3. If the web site has a high rate of change it may be helpful to take periodic snapshots of the entire web site

(See "6.3 Technical Guidelines for Scheduling and Managing Web Sites" for more information)

6.2.2 STATIC WEB SITES WITH FORM-BASED INTERACTIVITY

[This section does not refer to web based systems that facilitate a financial transaction between government and client(s).]

Many web sites utilize forms and "back end" information systems. These sites consist primarily of static html pages with simple form based interactivity. There are essentially two types of form-based interactive sites:

➢ Sites that collect information such as comments and requests from visitors, and
➢ Sites that provide a search interface for back-end information resources

Agencies should manage these sites as static web sites, but should take the additional steps of scheduling and managing:

➢ The "back end" information system
➢ The form itself
➢ The human readable source code of the script or program which enables the forms functionality (again, not necessarily scheduled, but "managed")

SCHEDULING

Evaluate each of these additional three elements of the web page and determine the appropriate retention period for each. Create and submit retention schedule.

IMPLEMENTING THE RETENTION SCHEDULE REQUIREMENTS

1. Work with webmasters and content creators to create procedures for updating forms and scripts and for maintaining deleted files for the appropriate retention period.
2. Include appropriate metadata in files that includes the date the page was made live; the date the page was removed from the web; and the proposed disposal date.
3. Move updated and deleted files to alternative storage medium for the length of the retention period.

(See "6.3 Technical Guidelines for Scheduling and Managing Web Sites" for more information)

6.2.3 WEB SITES BASED ON DYNAMIC DATA ACCESS

Web sites are sometimes used as front ends, or user interfaces, for accessing an agency's database(s). Site users search prepared lists or put together their own searches that, in turn, query the content of a database. The information returned from these queries is displayed as an HTML document to the user.

In many cases, documents exist as objects in a database. Each document will have its own unique identifier, usually reflected in the URL. This means that a user can bookmark the particular document and return to it later without reconstructing the original search query (provided the document has not been deleted from the database).

Even if the site's main or top level pages are static, dynamic data access web sites raise some additional issues for records management, including:

> Not all users "see" the same web site. At designated levels, the pages displayed on users' browsers are based on what they ask for, therefore user queries are an integral part of generating the web site and may need to be captured.
> Information contained in databases behind the site may be continually changing.

SCHEDULING

Scheduling these types of web sites can be complex. Agencies should review the section, 6.3 "Technical Guidelines for Scheduling and Managing Web Sites" and evaluate which method(s) are appropriate.

6.2.4 DYNAMICALLY GENERATED WEB SITES

An increasing number of web sites are being built which generate all of the pages "on the fly." This means that the component parts of each individual page—its content, structure, and presentation—are generated dynamically using a combination of databases and style sheets based on:

> A stored set of user preferences
> A stored set of access profiles
> A user query and/or
> The capabilities of the user's browser

In these situations, the web site does not exist in any single or easily capturable form. Each user sees a different "site" based on their stored preferences and access rights, current needs, and the capabilities or limitations of the technology they are using.

Although the end result for the user might be a set of static pages, the processes that build the pages involve the use of a number of software tools. This is the point at which web sites become more like software applications than electronic publications. Agencies need to consider how to manage and archive dynamically generated web resources in a fully functional state.

SCHEDULING

Scheduling these types of web sites can be complex. Agencies should review the section, "Technical Guidelines for Scheduling and Managing Web Sites" and evaluate which method(s) are appropriate.

6.3 TECHNICAL GUIDELINES FOR SCHEDULING AND MANAGING WEB SITES

There are essentially three overall strategies for scheduling and managing complex web sites:

➢ The object driven approach,
➢ The event driven approach, and
➢ Maintaining an online web site reproduction

Each of these approaches will be discussed below. However, the options suggested here are not mutually exclusive. Agencies should pursue a combination of strategies designed to fit their circumstances and environment.

OBJECT BASED VS. EVENT BASED APPROACHES

The major issues dynamic sites raise is the need to choose whether to use an object based or an event based approach to managing records of web resources and activities. That is, an agency needs to determine whether it wishes to focus on scheduling and managing records of:

➢ The objects that comprise the content of the site at any given time; or
➢ The individual transactions (events) between clients (users) and servers agencies)

It is important to note that the options suggested here are not mutually exclusive. In most cases agencies should pursue a combination of strategies designed to fit their circumstances and environment.

6.3.1 OBJECT-DRIVEN APPROACH

This approach concentrates on scheduling and managing the 'objects' that constitute or are made available via a web site. Adopting an object-driven approach for web sites that primarily provide transactional services may be futile.

In such circumstances, an event-driven approach may be more appropriate.

IMPLEMENTING THE OBJECT-DRIVEN APPROACH

This approach can be implemented in different ways, depending on your environment. These implementation strategies include:

Object Approach Implementation Strategy 1

Managing records of a dynamic site using the object-driven approach would involve keeping track of changes to enable its full reconstruction at any given date. This would require the capture, storage, and scheduling of:

- User profiles
- Style sheets
- Search engine
- Scripts and programs
- Regular snapshots of the database
- Database transaction logs

Object Approach Implementation Strategy 2

The object approach could entail taking periodic snapshots of collections of web resources in combination with tracking changes to the site and logging transaction details. The snapshots and tracking and transaction information should be included on agency retention schedules according to the content of the web resource(s).

Procedures for snapshots, tracking changes and logging transaction details are included as Appendix A, Object Driven Approach Implementation Strategy 2, to these Guidelines.

Object Approach Implementation Strategy 3

Alternatively, objects or individual web resources could be separately captured and managed in association with metadata that described the relationship between specified versions of the object and its unique URI. This approach focuses on managing the data objects and associated metadata instead of attempting to preserve entire systems that support web resources.

Procedures for managing objects separately are included as *Appendix B, Object Driven Approach Implementation Strategy 3*, to these Guidelines.

6.3.2 EVENT-DRIVEN APPROACH

This approach focuses of capturing 'events' or transactions that occur between the web site and the user.

This approach is most suited when a dynamically generated site is database-driven and relies on stored user profiles, search mechanisms, SQL-HTML translation scripts, and other programs to enable functionality. For these types of sites, it may be feasible to capture and schedule 'events'—single transactions between web site and user—rather than the objects that comprise the site at the time of the transaction. An event driven approach would involve capturing:

> ➢ Date and time of event
> ➢ IP or domain address of the user
> ➢ User profile
> ➢ Query or other action performed
> ➢ The resource served to the user with relevant metadata attached

Any web-enabled service or transaction facility provided by an agency will both generate and be made possible by records. In the absence of a record, there is no evidence of the transaction having occurred. In the absence of legally sustainable evidence of a transaction having occurred, the transaction may be repudiated and/or deemed by a court of law to have not taken place. It is therefore essential that agencies capture, schedule and manage full and accurate records of web-based transactions that can guarantee the authenticity, reliability and accessibility of the records.

Procedures for the event driven approach are include as Appendix C - Event Driven Approach Implementation Strategy, to these Guidelines.

6.3.3 MAINTAINING AN ONLINE WEB SITE REPRODUCTION

An online web site reproduction is intended to replicate, at the time of posting, all material posted to the active website. Unlike an active website, the online reproduction must capture past as well as present postings throughout their retention period. It is essentially an online "archive" of all the content ever included on the web site.

The main advantages of this option include:

> ➢ Facilitating maximum functionality of archived postings;
> ➢ Providing greater accessibility to archived postings, instead of reconstructing sites from an offline storage medium;
> ➢ Offering a more appropriate option for complex sites that are more than simple static electronic publications; and
> ➢ Enabling the reconstruction of the site at any point in time rather than only when a snapshot is created and therefore providing greater accountability for the agency.

Where an online reproduction is carefully planned and responsibilities are documented and assigned, there are few deficiencies associated with this option. Costs and other practical considerations associated with establishing and maintaining such a server will need to be balanced against benefits. In particular, maintaining all past and current website postings may require large amounts of storage space, with negative cost implications for this strategy.

Procedures for creating an online reproduction are included as Appendix D - Establishing an Online Reproduction of a Web Site Implementation Strategy, to these Guidelines.

CONCLUSIONS

The most appropriate records management strategy is likely to use a combination of approaches. For example, taking a regular snapshot of the site, maintaining a log of each change made to the website and maintaining a log of transactions.

Careful consideration of the deficiencies and advantages of each approach, along with other practical considerations, such as feasibility and cost-benefit analysis, will determine choices. The rationale for selecting a particular approach should be documented.

6.4 SELECTION OF STORAGE MEDIA

Depending on retention requirements, agencies need to decide whether to capture and maintain web-based records on an offline or online storage medium. The size and complexity of the records is one of the determinants of the choice of storage media. Snapshots of sites and activity logs, for example, are likely to consume large amounts of storage space.

A second determinant is the desired speed of access. There is usually some delay in accessing records stored offline. Options for offline storage include optical disk or magnetic tape. In contrast, online storage provides instantaneous access in the form of a hard drive. However, instantaneous access is more expensive to maintain, especially if the agency is storing large quantities of data.

Below is a brief description of four widely used storage media.

CD-ROM (compact disk—read only memory): Optical disk technology is capable of storing large amounts of data that can be read but not altered. CD-ROMs all conform to size and format standards and are well suited for color, large software applications, graphics, sound and video. CD-ROM technology adheres to ISO 9660 that covers both the physical layout of the disk and the format of the recorded information.

CD-R (compact disk—recordable): Based on WORM (write once read many) technology, a CD-R can store large amounts of data. CD-R technology also adheres to ISO 9660. CD-R drives have been improved to enable multi-session recording (that is, data can be added over time). Standard error checking techniques should be used to assess the quality of the blank discs being used for storage.

Magnetic tape: A magnetically coated strip of plastic on which data can be encoded, magnetic tape provides relatively inexpensive and large storage capacities. Because tapes do not allow random access to data, access time is slower on tape than on disks. Tapes are available in a range of sizes and formats.

Magnetic hard disk: A hard disk (as opposed to floppy disk) is a magnetic disk that can store large quantities of data. However, hard disk storage is more expensive than other storage media.

SECTION 7.0 - WEB-BASED PUBLICATIONS:

The State Library of Ohio is participating in a project to build a sustainable digital archive with the ability to track, preserve and provide long-term access to web-based publications that exist only in electronic form. For more information on this project, please look at the project web page at http://winslo.state.oh.us/govinfo/pep/pepintro.html. *[Please note that as of 10/15/2010 this is a dead link]*

In addition to making agency publications more accessible to the public, the State Library of Ohio will create preservation information and preserve the publication digitally in an electronic archive. In order to insure that your publications can be archived, we ask you to do the following:

➢ Create Dublin Core metadata to describe your publication
➢ Insert the metadata in publications on your website
➢ Notify us that a publication has been added to your website

7.1 SELECTING WEB PUBLICATIONS WITH ENDURING HISTORICAL VALUE

The State Library will evaluate web publications for inclusion in the electronic archive based on the following criteria:

➢ The web-based resource must meet the State Library's definition of a publication
➢ The uniqueness and value of the information contained in the web site
➢ The file formats contained within the web site
➢ The technical requirements for the web site
➢ The resource must be created ONLY in electronic format

7.2 PROCEDURES FOR NOTIFYING STATE LIBRARY OF WEB BASED PUBLICATIONS

➢ Insert Dublin Core metadata in HTML (see section 7.3 Using Dublin Core Metadata)
➢ Once your agency has added the attributes (METATAGS) to your major Web publication, replace the previous version on your server
➢ Notify PROJECT staff that you have information at your site ready for access by emailing to govinfo@sloma.state.oh.us *[Is this a valid email?]*
➢ State Library staff will review your submitted publication and add it to the archive at OCLC. That information is then used to create a catalog record for use in the is also forwarded to the centralized catalog of OhioLINK. Your publication is also harvested and archived at OCLC.

7.3 USING DUBLIN CORE METADATA

Dublin Core Metadata consists of fifteen (15) elements.

Title: The name given to the sources by the author/creator or publisher.

Author/Creator: The person(s) primarily responsible for the intellectual content of the resource, eg, author of a document, and illustrator of a visual resource.

Subject/Keyword: The topic addressed by the resource being described or keywords that describe the content of the resource.

Description: A textual description of the content of the resource, including abstracts in the case of document-like objects or content descriptions in the case of visual resources.

Publisher: The agent or agency responsible for making the resource available.

Other Contributors: The person(s) (e.g. editors and transcribers) or organization(s) who have made significant intellectual contributions to the resource.

Date: The date the resource was available.

Resource Type: The genre or category of the resource, e.g. novel, poem, dictionary or manual, working paper, directory.

Format: The data format of the resource, e.g. Postscript, HTML, text file, JPEG image. Resource Identifier String or number used to uniquely identify the resource, eg. URL, ISBN.

Source: Resource(s), either print or electronic, from which the resource is derived.

Language: Language of the intellectual content of the resource.

Relation: Reference to a related resource, e.g. executive summary, statistics separate from full report.

Coverage: The spatial locations and temporal duration characteristic of the resource, e.g. a place name

Rights Management: The content of this element is intended to be a link [a URL (Uniform Resource Locator) or other suitable URI (Uniform Resource Identifier) as appropriate] to a copyright notice, a rights-management statement, or perhaps a server that would provide such information in a dynamic way.

EMBEDDING DUBLIN CORE METADATA IN WEB-BASED PUBLICATIONS

Dublin Core Metadata can be used within the <head></head> section of an HTML document in order to describe and provide better searching of the page. This is called embedded metadata.

An example of embedded metadata is included below:

<META NAME="dc.title" content="Preserving Electronic Publications-Webmaster Guide">
<META NAME="dc.creator"content="State Library of Ohio">
<META NAME="dc.description" content="A guide targeted at state of Ohio agency webmasters and content creators responsible for publishing born digital web publications on their agency websites. Encourages use of Dublin Core Metadata and need for preserving publications for future generations.">
<META NAME="dc.date" content="1 June 2002">

CRUCIAL ELEMENTS

The following Dublin Core elements are crucial for participation in the current effort to identify, evaluate and preserve web-based publications:

➢ Title - a name by which the resource is normally known
➢ Creator - an entity (person, organization, or service) primarily responsible for the
➢ content of the resource
➢ Date - creation of resource or revision of resource
➢ Rights - ex. "For internal use only" or "Available to the public"

DESIRED ELEMENTS

➢ Subject/Keyword - words that describe the publication
➢ Description - an account (abstract) of the content of the resource
➢ Publisher - an entity responsible for making the resource available
➢ Contributor - an entity responsible for making contributions to the content of the resource
➢ Source - a reference to a resource from which the present resource is derived
➢ Language - only if the language is not English
➢ Coverage - spatial location, temporal period, or jurisdiction (notice they describe both the contents of the item and where to get more information on the topic)

METADATA TEMPLATE

The State Library has a DC metadata generator available on its web site at http://winslo.state.oh.us/govinfo/pep/metagen.html. *[Please note that as of 10/15/2010 this*

is a dead link] The following template can also be used for including the metadata elements in web-based publications:

Required Fields

<meta name="DC.Title" content="Insert the title of your publication here">
<meta name="DC.Creator" content="Insert the name of the person or entity that wrote or created the publication">
<meta name="DC.Date" content= "Insert the date the publication was created or revised">
<meta name="DC.Rights" content="Text or identifier of a rights management statement">

Desired fields

<meta name="DC.Description" content="Insert a statement describing the publication here">
<meta name="DC.Subject" content="Insert relevant keywords to aid in retrieval">
<meta name="DC.Publisher" content="entity responsible for making the resource available ">
<meta name="DC.Contributor" content="entity responsible for making contributions to the content of the resource ">
<meta name="DC.Language" content="insert language">

SECTION 8 - PRESERVING WEB-BASED RECORDS WITH ENDURING HISTORICAL VALUE:

Since the State Library of Ohio will preserve discrete publications that state agencies publish on their web sites, the State Archives will preserve entire web sites or portions of web sites of state agencies, board and commissions that have enduring historical value.

The State Archives of Ohio will review web sites from state agencies, boards and commissions that are closing and consequently can no longer maintain their own web sites. If State Archives determines that the web site is of enduring historical value, the files will be transferred to State Archives for permanent preservation.

In some cases State Archives will consider reviewing and transferring web sites of non-defunct agencies. Agencies are encouraged to contact State Archives before beginning a major redesign of their web sites so that State Archives staff can determine if the web site should be preserved. If an agency submits a retention schedule for web-based resources, the State Archives of Ohio will review the retention schedule. If the State Archives wants to appraise or select the records, that will be reflected on the retention schedule.

State Archives will preserve most web-based records in electronic formats. If State Archives determines that there is a danger of losing the information in the electronic format(s), they may choose to migrate the electronic information or to reformat the files to an analog format (if possible).

8.1 SELECTING WEB-BASED RECORDS WITH ENDURING HISTORICAL VALUE

State Archives will evaluate web sites based on the following criteria:

➢ The uniqueness and value of the information contained in the web site
➢ The file formats contained within the web site
➢ The technical requirements for the web site

8.2 REQUIREMENTS FOR TRANSFERRING WEB-BASED RECORDS WITH ENDURING HISTORICAL VALUE TO THE STATE ARCHIVES

➢ The web site files are included on an approved retention schedule
➢ State Archives has evaluated the web-based records and agreed to include them in the archives.
➢ State Archives will accept web site files that have been saved on either CDROM or 3.5 inch diskette.
➢ Before sending the files to State Archives, please make sure to check all links and scripts to make sure that they work.
➢ If changes to the original web page(s) have to be made in order to ensure usability, please record the changes that were made.
➢ State Archives can currently support the following file formats: HTML, TIFF, PDF, TXT, JPEG, GIF. If you have questions regarding file format issues, contact the State Archives.
➢ After evaluating the web site, State Archives staff will provide the agency with a list of metadata necessary to facilitate the transfer of the web site.

SECTION 9: BIBLIOGRAPHY

Archiving Web Resources: A policy for Keeping Records of Web-based Activity in the Commonwealth Government. National Archives of Australia. January 2001.

Archiving Web Resources: Guidelines for Keeping Records of Web-based Activity in the Commonwealth Government. National Archives of Australia. March 2001.

Preserving Electronic Publications: Webmaster Guide. State Library of Ohio. June 1, 2002.

APPENDIX A - OBJECT DRIVEN APPROACH IMPLEMENTATION STRATEGY 2

SNAPSHOTS

A snapshot usually involves creating a full and accurate record copy of an agency's public web resources at a particular point in time. If an agency decides to create period snapshots of a web site, the snapshots should be scheduled on a retention schedule. The snapshot should be maintained for the length of the retention period.

When taking snapshots of collections of web resources, it is desirable to ensure (as far as possible) the continuing processability of the website and its component pages. This means that agencies should try to retain the capability to replicate the content, layout and functionality of the site across technological platforms without loss of data integrity.

This strategy is particularly useful for static resources or collections of static objects that are essentially an agency's electronic publication(s).

A snapshot is an object-driven approach and should not be used to keep records of highly interactive dynamic sites or resources that are databases or transactional services. A deficiency of this approach is that a snapshot only provides a picture of a website at a particular point in time. If snapshots are captured in the absence of other records of web-based activity, it will be impossible to reconstruct the site together with its functionality at any other point in time. Since this method does not enable the agency to determine exactly when particular web resources were available, agencies that use the snapshots strategy should also create and maintain logs of changes made to web resources between snapshots.

PROCEDURES FOR CREATING AND CAPTURING A SNAPSHOT

A snapshot should include all aspects of the website to ensure that a fully functional site can be reconstructed. For example, the snapshot should also include scripts, programs, plug-ins and browser software, that is, all components that make the snapshot fully functional. The snapshot should be captured with sufficient descriptive metadata.

It may be necessary to make some modifications once the snapshot is created. For example, a CGI script for site counters will need to be disabled. If site counters are not disabled, there will be no accurate or authentic record of the number of visitors to the site at the time the snapshot was created. In effect, the record is no longer a snapshot of the site.

RESPONSIBILITIES

The main responsibilities to assign include:

➢ Determining if this is an appropriate option and whether it should be supplemented by other record management strategies;

➢ Determining how frequently copies of web resources should be created;

➢ Creating the snapshot;

➢ Capturing and maintaining of sufficient metadata for the length of the retention period;

➢ Selecting an appropriate storage medium and undertaking data management tasks and quality control check.

Website administrators or information technology staff may already carry out the task of creating 'back-ups' of the website as part of normal data management activities. However, because these back-up copies are created for the purpose of data management activities, they are usually overwritten regularly with more recent versions, or deleted. They are not captured or maintained for record management purposes. To be used as a viable records management strategy, it is necessary to intervene and establish processes and procedures to ensure that snapshots are created, captured and maintained over time for as long as required. The above responsibilities represent the minimum list of responsibilities that should be documented in agency's procedures and assigned to records management practitioners, website administrators and information technology staff.

TRACKING CHANGES

This strategy involves tracking changes to the web resources over time and creating a log of changes or activity. The activity log needs to be maintained to satisfy requirements for accessibility for as long as needed. Used in combination with snapshots of the web resources, this approach can be a reliable option for static sites.

The main problem arising from this option is the creation of insufficient metadata of the activity log, resulting in the inability to interpret the log over time. It is vital that metadata requirements are specified and sufficient metadata is captured.

Procedures for creating and capturing activity logs

Suggested data elements that can be captured in an activity log include:

➢ Title or name of posting;

➢ Version number;

➢ Author or content manager responsible for creating of the object;

➢ Links embedded in the posting;

➢ Date of initial posting;

➢ Date of modification;

➢ Date of replacement or withdrawal; and

➢ Disposal information.

This is not a complete list and agencies should review and adapt it to ensure their requirements are satisfied.

In the case of a static website, the log should capture changes to individual pages, documents or objects on the website. Changes to scripts, plug-ins, forms used to

present information etc. will also need to be captured as they will affect the functionality of the records.

It may be possible to use emerging web technologies to track changes. Web robots, spiders or crawlers are automated programs that visit sites for the purpose of indexing sites for searchengines. These programs may be useful for tracking changes.

Responsibilities

The main responsibilities to assign include:

➤ Determining the list of data elements that should be captured in an activity log;

➤ Establishing procedures and processes to ensure the activity log is created, updated and maintained over time;

➤ Capturing and maintaining the activity log, including the capture and

➤ maintenance of sufficient metadata;

➤ Selecting an appropriate storage medium and undertaking data management tasks; and

➤ Identifying preservation implications and ensuring the records are accessible for as long as required.

APPENDIX B - OBJECT DRIVEN APPROACH IMPLEMENTATION STRATEGY 3

PROCEDURES FOR MANAGING OBJECTS SEPARATELY

An agency should maintain a register or list of the URIs that have been made available on its public website and capture the data objects made available from these URIs. The relationship or association between the data objects and the URIs should be documented and maintained in a separate metadata store. The information about each URI and data object that should be captured includes:

- ➢ The absolute URI;
- ➢ The data object;
- ➢ The mime type of the data object;
- ➢ The start and end time of the association; and
- ➢ Possible relationships to other records that document the administrative processes by which the resource was authored and published.

This information enables the agency to accurately track web resources at any point in time. Although it is possible to read the metadata and then view each object individually, this method of archiving does not enable reconstruction of the website because the data objects have been separated from other components of the website.

RESPONSIBILITIES

The main responsibilities to assign include:

- ➢ Determining if this is an appropriate option and whether it should be supplemented by other record management strategies;
- ➢ Creating and managing the metadata store, including determining what metadata should be collected to document the relationship between each URI and data object (a relational database could be used to manage the metadata);
- ➢ Maintaining the data object store over time; and,
- ➢ Identifying preservation implications to ensure that web resources are accessible over time for as long as required (eg. storage media and migration requirements).

APPENDIX C - EVENT DRIVEN APPROACH IMPLEMENTATION STRATEGY

EVENT APPROACH IMPLEMENTATION STRATEGY

This event-driven strategy involves creating a log of site visitors, capturing the logs and any other records of web-enabled transactions together with sufficient metadata, and maintaining them as long as required. This option enables the capture and maintenance of evidence of site use, particularly any queries or transactions enabled by the site.

A problem may arise if appropriate log analyser software is not available. Although all web servers generate log files of server activity, raw log files are confusing and difficult to decipher.

A final issue to consider is privacy. Most logs of web site transactions will contain personal information about the user.

PROCEDURES FOR CREATING AND CAPTURING LOGS OF SITE VISITORS

A number of software packages are available that have the capacity to configure raw log files and produce comprehensive reports in the form of tables and graphs. The elements that can be logged include:

➢ Date and time;

➢ IP address or domain name;

➢ Pages visited;

➢ Actions performed;

➢ Queries made; and

➢ Web browser used.

Agencies will need to select the elements that satisfy their records management requirements.

RESPONSIBILITIES

The responsibilities for this option are the same as those listed in Tracking changes—creating and capturing activity logs.

APPENDIX D - ESTABLISHING AN ONLINE REPRODUCTION OF A WEB SITE IMPLEMENTATION STRATEGY

PROCEDURES FOR ESTABLISHING AN ONLINE REPRODUCTION OF A WEB SITE

The creation of an online web reproduction (a dedicated web server) should be carefully planned and designed to ensure both records management and system requirements are fully considered. Agencies should ensure that the necessary consultation occurs between records management practitioners, website administrators and IT staff. Modifications to website page counters will also need to be considered.

RESPONSIBILITIES

The main responsibilities to assign include:

➢ Determining whether this is an appropriate option, perhaps undertaking a cost-benefit analysis and comparisons with other options;

➢ Identifying system specifications;

➢ Setting up procedures to ensure that postings are captured;

➢ Ensuring sufficient metadata is captured;

➢ Identifying procedures and practices to ensure the system integrity of the online reproduction (eg data management tasks are identified and responsibility allocated);

➢ Identifying preservation implications and ensuring the web based-records are accessible for as long as required; and

➢ Implementing compliance programs to facilitate post-implementation review.

The above responsibilities represent the minimum list of responsibilities that should be documented in agencies' procedures and appropriately assigned. It is possible for agencies to add to and alter this list based on their specific requirements. The important point to emphasize in this approach is that there is a significant reliance on the early establishment of clear procedures. Consultation with various areas of the agency is essential for ensuring that the dedicated servers meet the requirements for the online web reproduction. To ensure the success of this approach, compliance with established procedures needs to be monitored.