

Preserving Presidential Library Websites

Amarnath Gupta
San Diego Supercomputer Center

Sponsored by

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
and
ADVANCED RESEARCH PROJECTS AGENCY
ITO INTELLIGENT METACOMPUTING TESTBED

ARPA Order D570
Issued by ESC/ENS under contract F19628-96-C-0020

January 18, 2001



SAN DIEGO SUPERCOMPUTER CENTER
TECHNICAL REPORT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.

Preserving Presidential Library Websites

A Case Study with the Franklin D. Roosevelt Library, Museum and Digital Archives

Amarnath Gupta

San Diego Supercomputer Center

La Jolla, CA 92093-0505

gupta@sdsc.edu

1. Introduction

As recorded at the website of the National Archives and Records Administration, the Presidential Library system formally began in 1939, when President Franklin Roosevelt donated his personal and Presidential papers to the Federal Government, based upon his belief that Presidential papers are an important part of the national heritage and should be accessible to the public. The Presidential Records Act of 1978 established that the Presidential records that document the constitutional, statutory, and ceremonial duties of the President are the property of the United States Government. After the President leaves office, the Archivist of the United States assumes custody of the records. The Act allowed for the continuation of Presidential libraries as the repository for Presidential records. The ten Presidential Libraries and the Nixon Presidential Materials Staff, maintain over 300 million pages of textual materials; 7 million photographs; 14.5 million feet of motion picture film; 83,000 hours of disc, audiotape, and videotape recordings; and 350,000 museum objects. The most important textual materials in each library are those created by the President and his staff in the course of performing the official duties. These historical materials form the substantive record of public policy in each administration. Other significant holdings include the personal papers and historical materials belonging to individuals associated with the President.

Recently the contents of the Presidential Libraries are being opened to the general public in the form of websites, where the original records and record series of the libraries are accessible as imaged documents, digitized photographs and videos. To aid the public's access to these large heterogeneous collections, textual documents have been indexed and transcribed into ASCII text, and annotations have been added to each collection of records to serve as a commentary that can help the users navigate through the website.

The goal of the current project is to determine how these valuable collections can be preserved within an arbitrary persistent store in a scalable, infrastructure independent manner. Each presidential web site imposes a knowledge structure that organizes digital objects to address issues related to important events that occurred during the presidency. An infrastructure independent representation for the web site must capture this knowledge structure.

2. The Research Questions

In this report we focus on the following set of closely related research questions:

- What does the notion of “infrastructure independence” mean in terms of creating a representation structure for any web-accessible digital library of archived objects?
- What would be an appropriate “infrastructure independent” representation for a Presidential Library?
- How automatable is the process of creating this representation starting from their current digital form of existence?
- How can this representation be made persistent in a manner that is independent of the technology that implements the persistence?

In the analysis presented here, we use the Franklin D. Roosevelt Library, Museum, and Digital Archives hosted at the Marist College as an example. We however believe that our analysis and observations are equally applicable to other digitally accessible Presidential Libraries. We present an approach that can be used to quantify the knowledge content of the web site, as well as information about individual digital objects.

3. Approach

The knowledge content required to support digital objects within Web sites can be quite complex. Web sites typically manage relationships that describe:

- Semantic meaning. Annotations are used to ascribe meaning to photographs and related events.
- Spatial correlations. Geographic Information Systems overlay maps, photographs, and images.
- Temporal correlations. Dates are associated with digital objects, and the temporal sequence of a series of messages is important.
- Conceptual spaces. The processes by which the digital objects were created needs to be described.

These types of knowledge can be organized in digital libraries. If it is possible to build a digital library representation that captures the structure of a web site, then the archiving of a web site can build upon other research efforts. The SDSC Technical Reports on “Knowledge-based Persistent Archives”[1] and “Towards Self-Validating Knowledge-Based Archives”[2] provide an overall description of knowledge management for digital libraries. The SDSC Technical Report “The Senate Legislative Activities collection (SLA)”[3], describes procedures for identifying the knowledge content within a collection, and for building archivable versions. The preferred approach for modeling a web site is to build the equivalent digital library. The ability to automate the process depends upon the amount of structure present within the web site.

Web sites are instances of digital libraries, with an underlying data structure and information organization. We consider the archiving of a web site in the context of the archiving of the associated digital library. This requires identifying:

1. Archival form to use for a knowledge-based description of a digital library
 - Temporal relationship description for the web site holdings
 - Type specific grammar selection for characterizing information content
 - Path signature semantics for accessing elements within the web site
 - Concept maps for representing domain knowledge (such as structure of government offices)
2. Archival form for modeling the web site
 - Rooted labeled graph
 - Database

4. Infrastructure Independence

Consider a library of *digital objects* with the following properties:

- Each object is described through metadata that can be represented in terms of a finite set of *attributes*.
- The objects may be *typed*, such that objects of a certain type will share a common set of attributes. The type could be enforced by the media of the object (audio, video etc.) or by the nature of the object (letter, telegram, memo etc.). The *type of an object can also be complex*, being defined in terms of sets and sequences of component sub-objects. Several letters in the Roosevelt Library are stored as multi-page documents, where each page has an image version and a transcribed text version. The types can possibly constitute a type hierarchy.
- Most *objects are organized in groups* by the values of some attribute. For the Roosevelt Library this is done by “file group” name (e.g., “Safe Files”, “Vatican Files”). The organization could be hierarchical so that within a given group some other attribute is used to form a subordinate group and so forth. In the Roosevelt Library “Safe Files” are organized by “box” numbers, and then further categorized by the nations/nation-groups that the objects correspond to.
- The library also contains a number of objects that are typed and have been placed in a collection but are *not associated* to any other object or group within the collection or to objects in other collection. This may reflect the fact that at any time not all associations between correlated objects may be explicitly documented in the library. The digitized photographs in the Roosevelt library are examples of apparently uncorrelated objects.

Aside from the digital objects that constitute the actual preserved entities, the web-based organization of the accessible materials impart an additional, and somewhat orthogonal ***annotation superstructure*** that consists of commentaries, descriptions, and a graph of links to the collections and objects of the digital library.

We believe that an infrastructure independent representation of a web-accessible digital library is a schema together with a set of index structures that satisfy each of the following conditions:

- *Every digital object and object collection in the original library is uniquely identifiable*
- *Every textual and non-textual object is represented in a default, lossless format*
- *The schema and the indices preserve sufficient information to reconstruct the original website from which the infrastructure independent representation was created*
- *The schema and indices can be augmented to create new object-groupings, inter-object associations and annotation superstructures to permit a different organization and access structure to the same original materials.*
- *The schema and indices can be automatically externalized through a universal encoding such that the physical technology supporting the web-based implementation of the library can be changed without affecting the content and structure of the library. The externalized representation should be information preserving and storable on a persistent medium.*

5. A Representation for a Presidential Library

We separate the problems of creating an infrastructure independent representation for the data objects in a Presidential Library and creating the same for the website that serves as the “annotation superstructure” of the data objects.

5.1 Representing the Data

A Presidential Library satisfies all the characteristics of a digital library outlined above. Additionally, in a Presidential Library:

- Almost every object is associated with a *timeline*, where the time is expressed at different *levels of granularity*. A letter has a date, a telegram may have a date and time, while a photograph may have a year or have a *named event* that can be associated with an interval such as 1939–44. Moreover, a specific digital object may have multiple references to a timeline. For example, a telegram may have a date and time of delivery, a date and time of receipt, and a narration referring to specific dates and times of events.
- A large subclass of objects shares a *common namespace* that consists of names of people, places, events and organizations. It is conceivable that with the help of the Presidential Librarian and staff, these names may be organized into a *semantic ontology* where the relationships between different elements of the namespace are known from auxiliary historical information.
- Some *inter-object relationships are typed*. For example, the Court Reform introduced in 1937 produced strong media reactions in the form of cartoons. In addition to the description of the cartoon as a digital object, its relationship with the event called “court reform” can be

expressed with the type “media reaction” as noted in the associated description in the web pages.

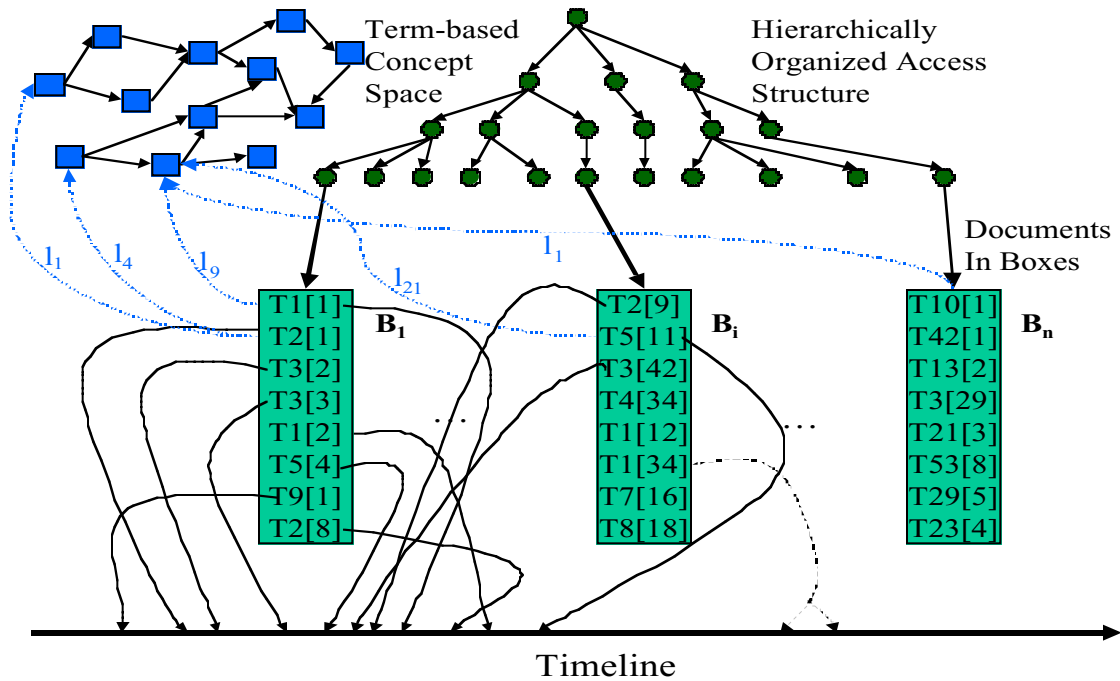


Figure 1. Information structure for mapping temporal data to concept spaces.

Figure 1 shows a schematic that illustrates the information structure we described. The blocks B_1 through B_n represent the record series in the library. Within each series, there are elements of different types each with multiple instances (T5[4] represents the 4th instance of the type T5). Each instance maps to an absolute timeline either as a point event (the solid line from B_1) or as an interval event (the dashed line from B_i). The hierarchical access structure pointing to the blocks represent the initial organization of the digital objects. Each node in the hierarchy represents a named “container” object that houses the next level of containers. Thus “Box 54” in the “Vatican Files” is a node in this diagram. The directed graph over the rectangular structures represent the names of higher-level terms or concepts interconnected through labeled links (labels not shown in the figure). We call this structure the *concept map* for the library, and consider it crucial to create a “knowledge-based archive” of digital libraries. As shown, multiple objects from different series might relate to the same concepts. The labels on the links connecting the objects and the concepts represent relation types as explained in the previous paragraph. We believe it is possible to construct a finite, albeit large, set of labels to describe these relationships. In future research we would consider methods to identify which data objects should be related to the concept space and the different classes of mappings between data elements and concepts (i.e., some possible classification over the set of labels shown).

Since the objects in each box are typed, and in many cases have textual transcription, each transcription may be expanded using a type specific grammar. Thus a letter can be decomposed into items such as the addressee's name and address, the dates referenced by the letter, the sender's name and title, the reference line describing the subject, the paragraphs, and possibly to sentences, each fragment being tagged suitably from a namespace commonly agreed among Presidential Historians and Librarians. As an example, assuming the root of the objects in the library is called "ORoot", a telegram in the Roosevelt Library with the access path *ORoot/Safe_Files/Box3/Great_Britain/* may be decomposed as follows:

```

<telegram object_id="3">
  <line count="1">AS</>
  <directive>
    <line count="2">THIS TELEGRAM MUST BE</>
    <line count="3">CLOSELY PARAPHRASED BE-</>
    <line count="4">FORE BEING COMMUNICATED</>
    <line count="5">TO ANYONE. (SC)</>
  </directive>
  <place_of_sender>
    <line count="6">LONDON</>
  </place_of_sender>
  <date_sent>
    <line count="7">DATED
      <month>DECEMBER</>
      <date> 6</>,
      <year>1940</>
    </>
  </date_sent>
  <date_received>
    <line count="8">REC'D
      <time> 9:10 A.M.,</>
      <date>7TH</>
    </>
  </date_received>
  <addressee>
    <name>
      <line count="9">SECRETARY OF STATE, </>
    </>
    <place>
      <line count="10">WASHINGTON.</>
    </>
  </addressee>
  <date_creation>
    <line count="11">3984,
    <date>DECEMBER 6, </>
    <time>MIDNIGHT.</>
  </date_creation>
  <directive type="confidentiality">
    <line count="12">STRICLTLY CONFIDENTIAL FOR THE SECRETARY AND THE UNDER </>
    <line count="13">SECRETARY AND FOR TRUITT MARITIME COMMISSION.</>
  </directive>
</date_information>

```

```

        <line count="14">MY 3965,
            <date>DECEMBER 5TH.</>
        </>
    </date_information>
    <body>
        <line count="15">THE SHIPPING SITUATION IS ... </>
        ...
    </body>
    ...
</telegram>

```

The XML syntax of the representation is the externalized version of an internal representation that can be stored in a modern object-relational or XML database. This would allow use of any XML-based query language devised to find information items by their *path signatures*. Thus, in the Roosevelt Library, the path *ORoot/Safe_Files/Box3/Great_Britain/telegram/directive* refers to the two directives in the object. Note in the representation that the actual line counts are used to reconstruct the digital object and are embedded within more descriptive semantic tags that capture the content of a generic telegram message. This example illustrates that a single digital object can have multiple temporal references with different time-granularities. Also note that several pieces of text are uninterpreted (e.g., “MY 3965”, line 14) and can be tagged only with domain-specific rules.

The information representation is supplemented by a concept map that indexes, for example, the addressee “secretary of state”, to an individual who fulfilled the role at the time the telegram was sent. The same concept map may contain the structure of the government to the extent reflected in the material covered by the digital objects in the library. In the same token it should have partial structures of the governments of other countries with whom correspondences are recorded in the library. A continuing research topic is the development of appropriate concept maps and illustrations of their use as finding aids.

Additionally, a *temporal index* should be used to allow a user to navigate through the library based on time independently of the initial organization of the document. The temporal index would be able to represent both point and interval representations of time and support simple temporal operations to locate objects. The index can be constructed from the data and externalized as an XML document. In future research, we shall also explore the possibility of constructing and using *spatial indices* to represent and reason over spatial data. Such data may include both spatial references, such as information reporting the deployment and movement of troops, as well as geographic entities like digitized maps that have been stored in the library.

We also believe that for the textual component of the library, *term indexes* need to be used to map terms to XML path constants to localize the term occurrences inside documents. The term index will help to manage multiple occurrences of or references to the same object in the library. This will aid a future researcher to explore potential regroupings that will facilitate the creation of a new structure

over the existing digital objects. We are currently investigating how such an index should be externalized for persistent storage.

5.2 Representing the Website

It has been amply demonstrated by recent Computer research dealing with website modeling, management, querying and restructuring, that websites are best modeled as rooted labeled graphs. For example, in the Strudel project from AT&T Labs a site graph is modeled as a labeled, directed graph, which contains *objects* and *collections*. Objects are connected by directed edges labeled with string-valued *attributes*. Objects are either internal nodes, identified by a unique object identifier, or are atomic values, such as integers, strings, and files. The model is externalized in XML using the following DTD.

```
<?xml encoding="US-ASCII"?>
<!ELEMENT STRUDEL (collections,objects)>
<!ELEMENT collections (collection*)>
<!ELEMENT collection (member*)>
<!ATTLIST collection name ID #REQUIRED>
<!ELEMENT member EMPTY >
<!ATTLIST member IDREF IDREF #REQUIRED>
<!ELEMENT objects (object*)>
<!ELEMENT object ANY >
<!ATTLIST object ID ID #REQUIRED>
<!ATTLIST object type CDATA #IMPLICIT>
```

Strudel supports several atomic types that commonly appear in Web pages e.g., URLs, and PostScript, ASCII, image, and HTML files. Objects are grouped into named collections, which are used in queries. Objects may belong to multiple collections, and objects in the same collection may have different representations. Data sources are either *tuple-stream data* sources, e.g., relational databases, shell scripts, text files, or *graph-structured data* sources, i.e., bibliographies, XML documents, and graphs that conform to Strudel's data model. We point out that since HTML and scripts together control the presentation of the pages of a site, Strudel's model is general enough to capture the presentation elements of a website. In addition, since Strudel's query language StruQL (or any other query language with the same functionality) allows restructuring of information, it inherently supports the creation of an alternate presentation structure to be defined as a "restructuring query" on the original presentation.

In order to adapt Strudel's model to a Presidential Library website, we take the following steps:

1. We treat *WRoot*, the root of the web site to be the root of a graph that is distinct from *ORoot*, the root of the hierarchical organizational structure leading to the data objects
2. For all references to internal digital objects, we replace URL references by *path expressions* leading to the referred object as shown in earlier examples.
3. All references that point to objects at external websites, should store, in addition to the external URL, the following:

- the physical location of the object (e.g., the name and address of the library)
- any available reference to the physical location of the object within the site (e.g., accession number of the object at the site)
- any metadata about the structure and format of the object (e.g., “600 × 480 jpeg image”), if available

We shall explore in future research if it is possible to define a minimal set of attributes for external references.

4. If a site is modified to change the presentation, content or any portion of the annotation superstructure that defines the website content, it will be store as a restructuring view without destructively altering the structure of the originally preserved site. Note that an insertion to the content is not a destructive change.

We believe this model is an adequate first approximation to represent any website in an infrastructure independent form. We shall investigate whether this model fails to represent any aspect of a Presidential Library website.

6. Automating the Process

The crucial step in automating the process of preserving the website in an infrastructure independent manner is to recognize how inherently structured the website is. For some websites, typically produced from a database, there are only a few different kinds of pages, each having a very uniform structure, and any arbitrary page has very little variation from this template structure. These websites can be easily “wrapped”, often quite automatically. On the other hand, some websites produce pages dynamically, and the structure as well as the content of the page varies significantly based on the user’s input and other variables. For these sites, the process of automatic is very difficult if not impossible. Most websites fall somewhere in between. The Gulf War website is connected to many external documents and has many external links. Archiving the Gulf War website requires archiving the external context simultaneously. Presidential Library websites, having a more standardized page structure and being more self-contained, are easier to wrap.

To preserve a web site, our goal is to convert it into a database, and then use our previously developed methods to preserve the database. The first challenge of wrapping a web site into a database is to determine what is the corresponding logical data model that truly represents all the information of the website. This logical data model must then be converted to a traversal plan, so that a web-crawler can visit the pages, and extract information to suitably populate the logical model. The true challenge is in the construction of the archivable form and comes from the variety of content, which as pointed out earlier, can be converted in XML only using some heuristics, that will succeed for some instances and fail for others. Our goal in the next phase of research is to determine generic ways to transform the website into a semantically richer database.

7. Conclusion

In this report we presented our proposed infrastructure-independent representation of Presidential Library websites and the design considerations we have used. We are currently investigating techniques to wrap the content of this type of web site into a database.

Acknowledgement

This research has been sponsored by the National Archives and Records Administration and Advanced Research Projects Agency/ITO, "Intelligent Metacomputing Testbed", ARPA Order No. D570, issued by ESC/ENS under Contract #F19628-96-C-0020.

References

1. R. Moore, "Knowledge-based Persistent Archives," SDSC Technical Report 2001-07, January, 2001.
2. B. Ludaescher, R. Marciano, R. Moore, "Towards Self-Validating Knowledge-Based Archives," SDSC Technical Report 2001-01, January, 2001.
3. R. Marciano, B. Ludaescher, R. Moore, "The Senate Legislative Activities Collection (SLA)," SDSC Technical Report 2001-05, January, 2001.