Challenge in Archiving Web Resources

Contents

[On with the show]

Challenge in Archiving Web Resources

Tyng-Ruey Chuang
Institute of Information Science
Academia Sinica
Taipei 115, Taiwan

PNC 2003 Annual Conference and Joints Meetings

Disclaimer

- A preliminary analysis on the issues involved.
- Mostly personal thoughts.
- Derived mainly from experience in two Web-mapping projects.

The Web & Archives

- Web as media: The Web as a mean to store and access archived materials.
- Web as resources: The Web as a source of materials that are worthy of being archived.
- Some materials are available only on the Web (blog, newsgroup, etc.).
- google.com: A snapshot of currently available Web resources.
- archive.org: Snapshots of then available Web resources.
- Is archiving snap-shooting?

Several Issues in Archiving Web Resources

- Resource identification: Place, authorship, date, authenticity, etc.
- Preserving the context: Preserve resource's content, its presentation, and its context. What function did the resource serve then, and what is the purpose of archiving the resource now?
- Storage and retrieval: Problems of efficient, persistent, sustainable, and easy access to the archived resources.
- Coping with changes: Dealing with updated content, changing URLs, "the deep Web," obsolete technologies, new standards, etc.

Some Characteristics of Web Resources

- Content is often mingled with presentation, and appears in certain context.
- Resources are experienced with the help of user agents (browsers, plug-ins, readers, etc.).
- Resources are media-rich. They employee a multiple of technology standards.
- Interactive resources: One may need several rounds of interactions with a resource in order to gain a full experience.
- Dynamic content: Content is generated by demand and is used to serve different needs (w.r.t. browsers, languages, domains, for example).
- The sustainability of the resource itself may depend on user participation (e.g., slashdot, sourceforge, eBay).
- Systems/software infrastructures are the prerequisite for any Web resource deployment and experience.

Generated by XSLies

Why Archive Web Resources?

- To get a snapshot of the appearance of the resources.
- To preserve the experience of using the resources.
- To be able to recreate the functionalities of the resources.
- To be able to incorporate the resources into other resources.

Who Are the Archivists?

- Resource users: By means of Web spiders, site mirroring, etc. The "outsiders".
- Resource providers: By means of backup, replication, versioning, etc. The "insiders".
- Web architects: "Archivability" by system design! Examples: NNTP and RFC 1036 (for newsgroup), RSS (for blog), and even XML (for self-descriptive data in general).

Lessons from the Newsgroup and the Blog

- Specifications more important than implementations. Get precise data formats and clear system protocols first, software realizations (many of them!) will follow.
- Publishing as archiving.
- Make it easy to re-create the resource from its publications.
- If a resource can be easily re-distributed, it can be easily archived.
- Archivability by system design!

Our Previous Work on Web-mapping

- Taiwan Social Map: Online aggregation and visualization of census data using SVG, XML, and free software. (http://tsm.iis.sinica.edu.tw)
- Retrofitting SEF (a de facto topographic map exchange format in Taiwan) to GML (Geography Markup Language) and SVG (Scalable Vector Graphics).
- To re-create our Web-mapping function in other context, or to incorporate the resources to others', is increasing an issue.
- How do we archive our results so that they are easily re-usable to others?

Web-based GIS: An Archivability Case Study

- They are media-rich, interactive, dynamic Web resources.
- Content is mixed with presentation and is produced in a context. Often special client-side agents are needed to browse maps.
- Why archiving Web-based GIS?
- Why is it difficult to archive a Web-based GIS? Because it is almost impossible to recreate its functionality from a Web-based GIS experience.

Ending Remark

- The Web is a rich source of archival resources.
- Want to archive a Web resource so as to recreate its functionality.
- Web publishing (blog, newsgroup, etc.) is Web resource archiving.
- Need to understand how better to publish GIS resources.
- Archivability by system design.