Archiving: Ensuring Long-Term Accessibility & Usability of CMS Content

Peter Van Garderen, Artefactual Systems
Email: peter@artefactual.com
Blog: http://archivemati.ca

Open Source Content Management Systems
(OSCMS) Summit

Vancouver, Canada. Feb 7-9, 2006

Archiving

- In the CMS/blogging world refers to moving and organizing outdated posts & articles (usually according to some calendar-based taxonomy)
- Also refers to moving digital information to a specific type of file format (e.g. PDF-Archival) or storage media (e.g. DVD) for long-term storage.
- Archiving does not include the additional digital preservation activities and infrastructure that is required to ensure the long-term access and usability of the digital information

Digital Preservation Issues

1. Storage media instability and deterioration
2. Technology obsolescence and incompatibility (at the level of: hardware, system software, application software, data and file formats, storage media readers and drivers)
3. Lack of metadata which results in
   - the failure to locate information,
   - the inability to render and read the information, or
   - the inability to attribute meaning or value to the information due to the lack of contextual information
4. lack of clearly assigned responsibilities and resources for long-term preservation
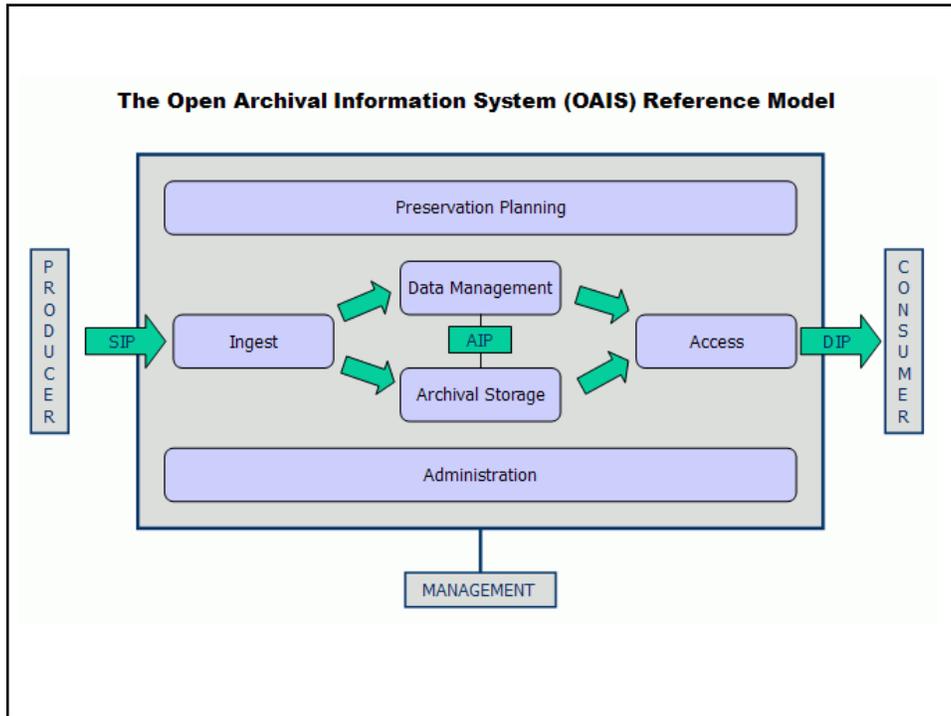
Digital Preservation Strategies

- **Technology Preservation**
  - preserving the technical environment that runs the system, including operating systems, original application software, media drives, and the like.
  - sometimes called the "computer museum" solution.
- **Analog Backups**
  - conversion of digital objects into analog form
  - microfilm is the safest bet for long-term preservation of information (only requires magnification and a simple light source to be read)
- **Migration**
  - periodic transfer of digital objects from one hardware, software and/or file format configuration to another, or from one generation of computer technology to a subsequent generation.
  - implementation of standardized file formats and normalizing or converting information to those formats.
- **Emulation**
  - Using emulators to allow programs or media that were designed for older, obsolete computing platforms to operate on current platforms
  - limited by the complex number of technical steps necessary to create functioning emulators, the administrative work to assemble specifications and documentation of systems to be emulated, and obtaining the intellectual property rights of relevant hardware and software.

Digital Preservation Considerations

- Long-Term Accessibility
- Usability / Functionality
- Authenticity
- Capacity
  - Technical
  - Financial
  - Staffing

Digital Archives

- a repository that stores one or more collections of digital information objects with the intention of providing **long-term access** to the information
  - (i.e. it incorporates a digital preservation strategy and management framework)

- 'Long-term' refers to a period of time which is long enough to be concerned about the impacts of changing technologies, including support for new media and data formats, and with a changing user community, on the information being held in a repository.

The Open Archival Information System (OAIS) Reference Model

CMS Components

- **Content**
  - database tuples
  - media files

- **Metadata**
  - CMS content types
  - Administrative data

- **Application**
  - Scripts, Code (core & plugins)
  - Themes
  - Servers

CMS Preservation Considerations

- Recordkeeping requirements?
- Just the content?
- Functionality?
  - "Deep Web"
- Look and Feel?

CMS Preservation Strategies

- Technology Preservation
  - Backup & restore procedures
  - Media Files?
- Snapshoting (e.g. Internet Archive)
  - Long-term access strategy (migration or emulation?)
- Syndication (e.g. RSS/Atom)
  - Is having just the content data sufficient?
- Retention Rules built into the CMS
  - Output to PDF-Archival?
- Preservation Format for CMS/Blogs?