Preserve Data Sets

Table of contents

- Activity Definition(s)
- Scholars' Stories (scenarios)
- Tools (examples)
- Related Collections/Content (examples)
- Applicable Standards or Standards Bodies
- Notes, comments, related activities, concerns

go back to Activity Definitions page

Preserve Data Sets

Keywords: Preserve

Activity Definition(s)

Data sets are interpreted for this activity definition to be more-or-less quantitative or tabular in nature, or the raw results of research investigation (e.g., field notes depicting where individual archeological objects were found in an excavation site, or measurements of such objects), or primarily machine-readable and interpretable content. (Other activity definitions deal with preservation of other content types that are more human-interactive than data sets. Note there is some overlap here with the activity Preserve a visualization or a model, which mentions GIS data sets specifically.)

Some examples might be historic census records, or measurements of visible characteristics of artistic or cultural objects, or counts and occurrences of stylistic units and patterns found within a corpus of text or within a musical work. Much work may go into generating such data sets, and the results may be preserved for others to access and build upon.

- Determine scope of preservation - entire data set, or key excerpts for future use?
- Identify the necessary components to be captured and how to capture them.
- Obtain copies of all the necessary components.
  - Data and metadata
  - Methods and procedures used
  - Other supporting documentation.
- Assemble, edit and verify the components as needed.
- Migrate components to more durable formats as needed.
- Preserve materials in a digital storage facility.

Scholars' Stories (scenarios)

- A prosthetic reading of Shakespeare's genres - University of Wisconsin-Madison deals with quantitative data and its analysis.

Tools (examples)

----Optional: examples of tools that perform some part or all of the defined activity, ideally with links to relevant project or vendor web sites

<table>
<thead>
<tr>
<th>Tool name</th>
<th>What it does</th>
<th>Relevant links</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATAstor</td>
<td>DATAstor is a &quot;proto-demonstrator&quot; within the Tools &amp; Content Partners working group, proposed by John Laudun. The contents of DATAstor will be materials collected by humanities field researchers, which can include, but is not limited to, the following: oral texts: oral histories, myths, legends, anecdotes, jokes, songs, proverbs, dites, etc. material artifacts: houses, tools, boats, etc. performances: festivals, rituals, marketplace interactions</td>
<td><a href="https://wiki.projectbamboo.org/display/BPUB/DATAstor">https://wiki.projectbamboo.org/display/BPUB/DATAstor</a></td>
</tr>
<tr>
<td>Pronom</td>
<td>File format registry for digital content review</td>
<td><a href="http://www.nationalarchives.gov.uk/pronom/">http://www.nationalarchives.gov.uk/pronom/</a></td>
</tr>
</tbody>
</table>
Related Collections/Content (examples)

--- Optional: examples of collections / digital content / digital resources that could be involved in part or all of the defined activity, with links to relevant repository or site where available

<table>
<thead>
<tr>
<th>Collection/content name</th>
<th>Collection/content description</th>
<th>Relevant links</th>
</tr>
</thead>
</table>

Applicable Standards or Standards Bodies

--- Optional: examples of standards or standards-bodies applicable to the defined activity

<table>
<thead>
<tr>
<th>Standard name / body</th>
<th>What it governs/regulated/standardizes - What it's for</th>
<th>Relevant links</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREMIS (PREservation Metadata Implementation Strategies)</td>
<td>Metadata standard for encoding preservation Information</td>
<td><a href="http://www.loc.gov/standards/premis/">http://www.loc.gov/standards/premis/</a></td>
</tr>
</tbody>
</table>

Notes, comments, related activities, concerns

Note 1. This activity was originally called “Store sets of data that comprise a research corpus.”

Note 2. The Perseus Project outlines three categories of information and access (see http://www.perseus.tufts.edu/hopper/research), and it seems the latter two categories are within the scope of humanities data sets:

**Human readable information:** digitized images of objects, places, inscriptions, and printed pages, geographic information, and other digital representations of objects and spaces. This layer of functionality allows us to call up information relevant to a longitude and latitude coordinate or a library call number. In this stage digital representations provide direct access to the physical senses of actual people in particular places and times. In some cases (such as high resolution, multi-spectral imaging), digital sources already provide better physical access than has ever been feasible when human beings had direct contact with the physical artifact.

**Machine actionable knowledge:** catalogue records, encyclopedia articles, lexicon entries, and other structured information sources. Physical access can serve our senses but provides no information about what we are encountering - in effect, physical access is like visiting a historical site about which we may know nothing and where any visible documentation is in a language that we cannot understand. Machine actionable knowledge allows us to retrieve information about what we are viewing. Thus, if we encounter a page from a Greek manuscript of Homer, we could at this stage find cleanly printed modern editions of the Greek, modern language translations, commentaries and other background information about the passage on that manuscript page. If we moved through a virtual Acropolis, we could retrieve background information about the buildings and the sculpture.

**Machine generated knowledge:** By analyzing existing information automated systems can produce new knowledge. Machine actionable knowledge allows, for example, us to look up a dictionary entry (e.g., *facio*, "to do, make") in a dictionary or to find pre-existing translations for a passage in Latin or Greek. Machine generated knowledge allows a machine to recognize that *fecisset* is a pluperfect subjunctive form of *facio* and to provide reading support where there is no pre-existing human translation. Such reading support might include full machine translation but also finer grained services such as word and phrase translation (e.g., recognizing whether *orationes* in a given passage is the object of a given verb), named entity identification (e.g., identifying Antonius in a given passage as a personal name and then as a reference to Antonius the triumvir).

Note 3. Here are a few examples of what might be considered humanities data sets, selected from the “NCSU Libraries web page of Social Science & Humanities Data Sets” (http://www.lib.ncsu.edu/data/socialscienceandhumsets.html):

American Religion Data Archive
“Includes data on churches and church membership, religious professionals, and religious groups (individuals, congregations and denominations).” Data with greatest detail is for U.S. but includes international surveys as well.

Archaeology Data Service

http://ads.ahds.ac.uk

A digital library providing access to archaeological archives and publications for Great Britain and Ireland.

Cultural Policy & the Arts National Data Archive (CPANDA) [Princeton University]

http://www.cpanda.org/

“An interactive digital archive of data on the arts and cultural policy in the U.S., available for research and statistical analysis, with data on artists, arts and cultural organizations, audiences, and funding for arts and culture.”

Pew Internet and American Life Project Data Sets

http://www.pewinternet.org/data.asp

“The Pew Internet & American Life Project will create and fund original, academic-quality research that explores the impact of the Internet on children, families, communities, the workplace, schools, health care and civic/political life. The Project aims to be an authoritative source for timely information on the Internet’s growth and societal impact, through research that is scrupulously impartial.” Poll data is made publicly available six months after release. Important source of U.S. public opinion data.

U.S. Bureau of the Census

The U.S. Census Bureau provides much of its data through American FactFinder. To retrieve smaller amounts of data, entry through the Data Sets (in the sidebar)--Detailed Tables links is strongly recommended. See the Data Download site for larger downloads from the major surveys. Some ACS data sets are only available through the ACS FTP web page. The Current Population Survey is only available through DataFerrett. County Business Patterns and other special topic data is available through the Subjects A to Z index from the Bureau’s home page and/or the CenStats databases. (The Agricultural Census has been administered by the USDA since 1992.) See our Census guide for more information, including how to access what historical data is available online (1980 and earlier). If you need assistance understanding or using Census data, contact the Data Services Librarians.

go back to Activity Definitions page