Dianne McCutcheon
Chief, Technical Services Division, Library Operations
NLM, NIH, HHS
mccutcd@nlm.nih.gov

Digital Archive Preservation and Sustainability Workshop
Baltimore MD, September 22, 2008

Digital Preservation and Sustainability:
Some Thoughts Based on Experiences at
the National Library of Medicine

Dianne McCutcheon
Chief, Technical Services Division, Library Operations
NLM, NIH, HHS
mccutcd@nlm.nih.gov
Overview

- NLM’s mission and its long-range planning goals

- Key aspects of digital preservation and sustainability as related to selected digital preservation initiatives of the NLM
NLM Statutory Functions
(NLM Act, 1956)

- Acquire and preserve library materials pertinent to medicine
- Organize the materials by appropriate cataloging, indexing, and bibliographical listing
- Disseminate the catalogs, indexes, and bibliographies
- Make the library materials available through loans, copying procedures, or otherwise*
- Provide reference and research assistance*

*NLM is national backup to other members of the National Network of Libraries of Medicine (MLAA, 1965)
Relevant NLM policy

- Preserve intellectual content if preserving both content and original format not feasible
- As first priority, preserve content of the journals indexed by NLM and its predecessors
- Encourage publication in permanent media to reduce future preservation costs
Types of Data

- Databases – e.g., bibliographic, genetics/microbiology, toxicology and environmental health, vocabulary (UMLS), clinical trials, etc.
- Text (with some multimedia) – published literature, Web documents, interactive publications
- Image data – e.g., video, historical pictures and films, Visible Human, 3-D structures
Goal 1: Seamless, Uninterrupted Access to Expanding Collections of Biomedical Data, Medical Knowledge, and Health Information

- Preserve NLM’s collections in highly usable forms and contribute to comprehensive strategies for preservation of biomedical information in the U.S. and worldwide

- Structure NLM’s electronic information services to promote scientific discovery and rapid retrieval…
Use digital technology to preserve and provide permanent access to additional parts of NLM collections

Assist in development of workable long-term national and worldwide strategies to guarantee preservation of key biomedical information – including geographically dispersed repositories

Pursue immediate opportunities, including agreements with interested publishers and international PMC partners
Long Range Planning
Promote Discovery

- Engage NLM’s research and service divisions in improving discovery and retrieval from existing NLM resources

- Conduct R&D in advanced information retrieval, synthesis and display
Digital Preservation and Sustainability

Key Concepts

- Preserving library data and scientific data not that different
- Collection and Curation
  - Selection
  - Standards – format and structures needed
- Access
  - Reduce barriers to use
  - Provide ways to link information
Digital Preservation and Sustainability

Key Concepts Cont’d.

◆ Storage and Migration
  ● Standards
  ● Community support

◆ Funding and Support
  ● Relying on single source increases risk
  ● Need visibility and use
Digital Sustainability Concepts in Selected NLM Digital Initiatives

- Published literature
- Genetic sequence data
- Archiving NLM’s web pages
### PubMed Central Journals — Tabbed List

Search by part or all of a journal name.

<table>
<thead>
<tr>
<th>Search this Journal</th>
<th>Title</th>
<th>Volumes in PMC</th>
<th>Free Access</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Infection and Immunity</td>
<td>v.72(6) Jun 2004</td>
<td>v. 1 1970</td>
</tr>
<tr>
<td></td>
<td>International Journal for Equity in Health</td>
<td>v.3(1) 2004</td>
<td>v. 1 2002</td>
</tr>
<tr>
<td></td>
<td>International Journal of Health Geographics</td>
<td>v.3(1) 2004</td>
<td>v. 1 2002</td>
</tr>
<tr>
<td></td>
<td>Journal of the American Medical Informatics Association</td>
<td>v.10(6) Nov 2003</td>
<td>v. 1 1994</td>
</tr>
</tbody>
</table>

- **New...** tab lists journals added to PMC in the past 60 days.
- **Title** links to a list of all issues of the journal in PMC.
- **Latest Volume** links to the most recent issue available in PMC.
- **Free Access** says how soon after publication the journal's articles are made free.
## PubMed Central Growth

<table>
<thead>
<tr>
<th></th>
<th>April 03</th>
<th>August 08</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journals in PMC</td>
<td>121</td>
<td>430</td>
<td>255%</td>
</tr>
<tr>
<td>Articles in PMC</td>
<td>84,637</td>
<td>1,799,910</td>
<td>2026%</td>
</tr>
<tr>
<td>Unique Users</td>
<td>273,207</td>
<td>5,0127,776</td>
<td>1776%</td>
</tr>
<tr>
<td>Articles Accessed</td>
<td>701,046</td>
<td>11,134,584</td>
<td>1488%</td>
</tr>
</tbody>
</table>
Introduction

The National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) created the Journal Archiving and Interchange Tag Suite with the intent of providing a common format by which publishers and archives can exchange journal content. The Suite provides a set of XML schema modules that define elements and attributes for describing the textual and graphical content of journal articles as well as some non-article material such as letters, editorials, and book and product reviews.

The Suite of Modules

The intent of this Tag Suite is to preserve the intellectual content of journals independent of the form in which that content was originally delivered. The Suite has been written as a set of XML schema modules, each of which is a separate physical file. No module is an entire schema by itself, but these modules can be combined into a number of different schemas.
UK PubMed Central: an international initiative

Based on PubMed Central (PMC), the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature, UK PubMed Central provides a stable, permanent, and free-to-access online digital archive of full-text, peer-reviewed research publications.

UK PubMed Central is part of a network of PMC International repositories. PubMed Central International is a collaborative effort between the U.S. National Library of Medicine (NLM), the publishers whose journal content makes up the Pubmed Central archive, and organisations in other countries that share NLM’s interest in archiving life sciences literature.

Through UK PubMed Central we aim to provide a freely accessible, UK-based archive of biomedical and health research findings.

As of July 2008, UK PubMed Central holds over 1.3 million full text articles.

Developing a resource for UK biomedical and health researchers

The ambition of UK PubMed Central is to become the information resource of choice for the UK biomedical and health research communities. We hope to achieve this by:
Gene Sequence Data

- International Nucleotide Sequence Consortium collaboration between 3 repositories of genetic data (European, U.S. and Japanese)
- NLM developed standards for exchange of data on a daily basis between the 3 repositories
- Led to integrated access to sequence, mapping, taxonomy, and structural data
- Ability to retrieve related sequences, structures and references
Four categories of permanence have been defined:
1) Permanent, unchanging content
2) Permanent, stable content
3) Permanent, dynamic content
4) Permanence not guaranteed
Fact Sheet

Unified Medical Language System®

Background:
In 1986, the National Library of Medicine (NLM), began a long term research and development project to build a Unified Medical Language System (UMLS®). The purpose of the UMLS is to aid the development of systems that help health professionals and researchers retrieve and integrate electronic biomedical information from a variety of sources and to make it easy for users to link disparate information systems, including

Last updated: 18 July 2003
First published: 01 January 1988
Permanence level: Permanent: Stable Content

Replaced by

U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894 Copyright, Privacy, Accessibility National Institutes of Health, Department of Health & Human Services
Summary

- Preserving digital information presents great challenges to both libraries and the scientific community

- Lessons learned by the library community about preserving collections can offer guidance in the challenges facing us all in preserving our digital heritage