2010

Building the Pandemic Influenza Digital Archive (PIDA) at the National Institutes of Health Library

James King

National Institutes of Health, james.king@nih.gov

Follow this and additional works at: http://jdc.jefferson.edu/scitechnews

Part of the Physical Sciences and Mathematics Commons

Let us know how access to this document benefits you

Building the Pandemic Influenza Digital Archive (PIDA) at the National Institutes of Health Library

By James King, National Institutes of Health Library, Office of Research Services, National Institutes of Health, Bethesda, MD 20892-1150, USA

Introduction

The NIH Library and the National Institute of Allergy and Infectious Diseases teamed up starting in 2009 to create a unique Web-based collaborative space to assist historical pandemic influenza researchers around the world. Built upon the innovative outreach of the library's Informationist program, this effort applies the concept of a Virtual Research Environment (VRE), using open source software, to enable global collaboration and expose a unique collection focused on historical flu pandemics.

The National Institutes of Health (NIH) is often referred to as the "crown jewel" of the Federal Government, serving as the "steward of medical and behavioral research for the Nation." With a sprawling 310-acre campus in Bethesda, MD, NIH comprises 27 Institutes and Centers and employs 18,000+ people, roughly half in scientific and clinical positions. NIH conducts translational bench-to-bedside health care research.

The NIH Library has a staff of 48 full-time employees and 15 contractors. The Library supplies research information to the residents of the NIH campus: intramural researchers who work in the laboratories and clinics, as well as grant administrators. The Library's services and collections are comparable in size and scope to a large academic biomedical library.
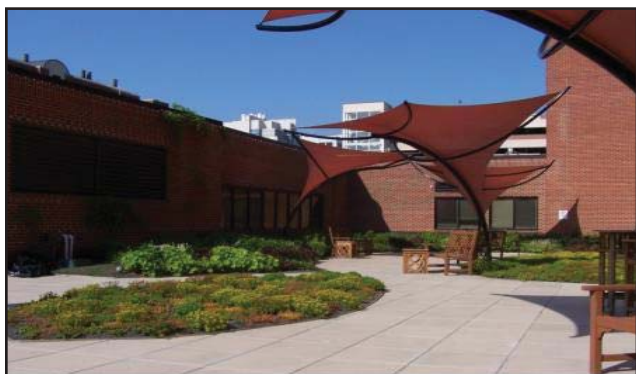
The NIH Library's Green Terrace

As surveys over the past decade have consistently demonstrated, the NIH Library is highly valued for its specialized activities and services. However, as in many research and academic institutions, the shift to a digital environment has forced fundamental changes in services and library space.

The NIH Library reacted to this shift by transforming the physical library into an information commons and also by developing Informationists—specialized librarians embedded in the researcher's context.

The Informationist Program, also known as the Librarian-in-Context service, was initiated to improve outreach and service to clinical research groups at NIH. This program, which began in 2001 with one Informationist at one Institute, has expanded over time to include 14 Informationists working with more than 40 groups in 16 institutes and centers. Having librarians embedded in specialized research groups allowed the Library to more easily integrate information services and resources into the workflow of NIH clinical and bench scientists and science administrators. Informationists contribute significantly to the scientific process, introducing a wider range of resources to the research activity, increasing scientists' satisfaction with their ability to answer questions, and enabling pursuit of answers to meet information needs. (Grefsheim, 2009) [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859271/]

Virtual Research Environments

One of the newest services being provided through the Informationist Program is the creation of custom databases for specific work projects. Initially, the Informationist delivered a literature search in the form of an Endnote file or an ASP/SQL based web page. This approach was functional but rudimentary and time intensive. A more rapid, streamlined approach was called for. In addition, several new projects required more than a good bibliographic database. They also needed strong collaboration tools to support the creation of communities around the topic.

As we researched a solution, Virtual Research

Environments (VRE) sounded like the perfect concept for us to focus on. VRE's were described by the UK-based JISC as "the tools and technologies needed by researchers to do their research, interact with other researchers (who may come from different disciplines, institutions or even countries) and to make use of resources and technical infrastructures available both locally and nationally." (JISC, 2010)

The Robertson Library at the University of Prince Edward Island has made an extensive effort to build VRE's while ensuring a robust preservation of the objects contained within a VRE using open source Fedora Commons digital repository software. Islandora [http://islandora.ca/] is an open source project that "combines the Drupal and Fedora software applications to create a robust digital asset management system that can be used for any requirement where collaboration and digital data stewardship, for the short and long term, are critical."

For an example closer to our medical research environment, we turned to a collaborative effort by the Massachusetts General Hospital and Harvard University to build a software toolkit called the Science Collaboration Framework (SCF) [http://sciencecollaboration.org/]. The SCF is also built on the open source Drupal software and is designed to foster the rapid development of web-based virtual team organizations for researchers in biomedicine. From their description, it "enables researchers to publish and discuss on-line content such as articles, news, and perspectives, and to provide shared semantic context for this content using established scientific vocabularies and automated text mining. SCF supports scientists in publishing, annotating, sharing and discussing content such as articles, perspectives, interviews and news items, as well as providing personal biographies, formal and informal bibliographies, and asserting research interests. SCF also supports shared databases of key research resources and private research workspaces." Two popular websites based upon the SCF are Stembook [http://www.stembook.org], "a comprehensive open-access collection of original, peer-reviewed chapters covering topics related to stem cell biology," and Parkinson's Disease Online [http://www.pdonlineresearch.org/], "a collaborative community for technical discussion and problem-solving in Parkinson's disease science."

NIH Library staff concluded that the concept of a Virtual Research Environment was not only viable, but could also be managed in a library setting and built using reusable components. We decided to move forward with Drupal [http://drupal.org] to build our Informationist-led development projects.
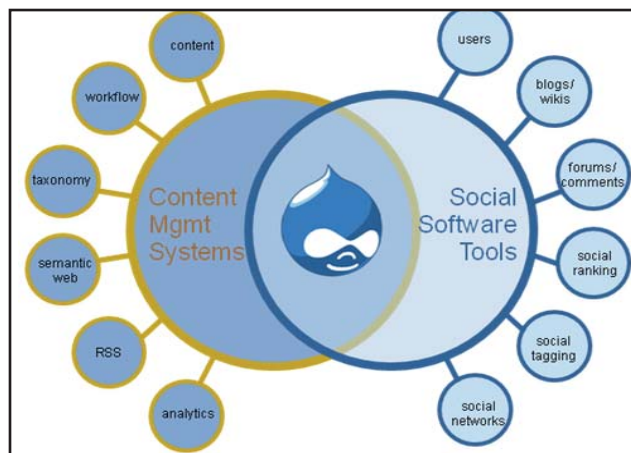
Drupal brings together features of CMS and Social Software Tools in a single platform

Drupal is a database-driven Web architecture specifically designed to serve as a content management system (CMS) and as a social publishing system. CMS gains a distinct advantage over previous iterations of Web site development by moving all of the content, scripts, images, and presentation styles into a single database rather than hosting in a myriad of individual files. Built in a modular fashion, Drupal has dramatically grown since its founding nearly a decade ago, boasting over 400,000 Web sites, 4,000+ modules, and a developer community of well over 350,000 members. Drupal offers many "core" features including account creation, rights management, syndication feed (RSS) creation, blog/wiki administration, and support for classic or fully interactive and collaborative Web sites. These core features can be expanded through the addition of modules created by organizations and individuals and made available at no charge. Additional modules being utilized in this instance include a module to support LDAP/Active Directory integration (LDAP), a bibliography module (Biblio) to manage large lists of publications, and a taxonomy vocabulary creation and management module (taxonomy). The NIH Library has partnered with Acquia [http://acquia.com], the commercial arm of the open source Drupal community, to obtain an expanded Drupal core, to ensure

access to technical support, and to take advantage of integration services.

Pandemic Influenza Digital Archive

In response to the health community's critical need for an accessible, centralized source for historical influenza data, the National Institutes of Health Library and the National Institute of Allergy and Infectious Diseases' (NIAID) Office of Communications and Government Relations are collaborating on the creation of the Pandemic Influenza Digital Archives (PIDA). The goals are to facilitate the ability of scientists and researchers, both within and outside NIH, to explore and respond to current issues and ideas and to acquire a deeper understanding of pandemic influenza.

NIAID conducts and supports basic and applied research to better understand, treat, and ultimately prevent infectious, immunologic, and allergic diseases. For more than 60 years, NIAID research has led to new therapies, vaccines, diagnostic tests, and other technologies that have improved the health of millions of people in the United States and around the world.

Over the past 30 years of his career, Dr. David Morens, Senior Advisor to the Director, has identified, collected, and compiled a core collection of research information focused on the epidemiology, etiology, diagnosis, and treatment of all pandemics and large scale epidemics, especially the 1918 pandemic influenza and influenza-related diseases. This collection presently serves his needs as a virologist with a special interest in pandemic influenzas. He and his colleagues have utilized data from documents in dozens of publications on the history of influenza and its significance for the future. Currently, Dr. Morens has a collection of more than 5,000 documents, including journal articles, statistical data, email correspondence, books and book chapters, bibliographies, reviews, and abstracts. Documents in this collection span the following topics:

- Bacteriology
- Clinical/Complications
- Enzootic Influenza
- Epidemiology and Events
- Etiologic Studies
- Experimental Infection
- World Regions
- Military Populations
- Civilian Populations
- Morbidity and Mortality Data
- Outbreaks (localized)
- Pathology
- Pathogenesis
- Public Health Disease Control
- Radiologic Findings
- Streptococcal Infections
- Treatment
- Tuberculosis
- Vaccinations

This collection serves as the starting point for a comprehensive and vital pandemic influenza digital archive. The goal is to facilitate the ability of scientists and researchers, both within and outside NIH, to explore and respond to current issues and ideas and to acquire a deeper understanding of pandemic influenza.

The NIH Library's initial plan for PIDA focused on creating a bibliographic database of the documents held by Dr. Morens by cataloging and richly tagging each item in the collection. Using Drupal to create a prototype, the NIH Library was able to show the value of Virtual Research Environments and power of collaboration built on top of this unique collection. NIAID eagerly agreed to switch gears and rescope the effort to reflect this broader goal—which was actually in line with Dr. Morens' ultimate goals for the project.
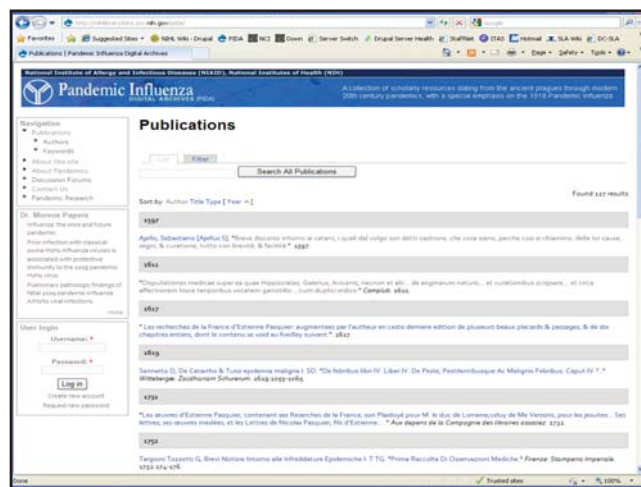
The PIDA Web site, built from the ground up to be a virtual collaboration space for historical influenza virologists, will showcase Dr. Morens' core collection. The prototype version of this Web site focuses on the 100 best papers in the collection, as identified by Dr. Morens. Chosen to represent the breadth and depth of the col-

lection, these initial documents will serve as the proving ground for key features of the site.

A key module of the PIDA site is the Drupal Biblio module, which manages lists of scholarly publications. Biblio can import and export various formats, including BibTex, EndNote, and XML, and can display citations in AMA, APA, Chicago, IEEE, MLA and Vancouver formats. This module is also able to interface with the Drupal Taxonomy module, which allows uploading of existing taxonomy vocabularies or the creation of custom vocabularies.

Basic records are being enhanced by a professional indexer to tag various aspects of each publication, including the date, geographical location, age, gender, ethnic group, weather conditions, type of disease, and local setting being described. We are following the National Library of Medicine's (NLM) cataloging rules and using NLM-maintained Medical Subject Headings (MeSH) as much as possible but we're also creating custom indexing standards and vocabularies based upon the needs of the collection and users. Rich and custom tagging of each publication will improve findability and provide data points for more advanced visualization of the collection, especially by global location and timeframe.

Since the Web site is being designed to support collaboration, registered users will be able to add custom tags to each record, vote for documents based upon the 5-Star scale, and add comments to each publication record.

The site is also being designed to support scholarship and reusability so registered users will be able to save a subset of publications into a public or private custom library. Publications that have been saved into a public library will be marked to show what custom sets they are reflected in, allowing for the community to build collections of documents that will benefit the community. For example, if a researcher studied the pulmonary effects of flu on children in the 1918 pandemic and marked all documents in the collection that were related, this would be of interest to a future researcher studying all effects of the flu on children.

The NIH Library is also working with vendors to perform focused literature searches of historical indexes (including Thomson-Reuters' Social Science Citation Index, the Global Health Archive, and ProQuest newspapers) to not only make the collection as complete as possible but to also introduce "copy cataloging" processes into the effort as much as possible.

A scanned copy of each article will be linked to each record and used by internal staff to enhance each record's index. Those articles that are outside copyright will ultimately be made available to NIH from the PIDA Web site. Since many of the older works are not in English, work has also begun to translate the bibliographic record for each publication into English.

PIDA is the first of many VRE's designed by the NIH Library based upon Drupal. Development of subsequent environments is being accelerated, by reusing common modules such as Biblio and repurposing common elements such as taxonomy vocabularies. Overall, this has been a positive experience and has opened new service opportunities for the NIH Library and its Informationist Program.
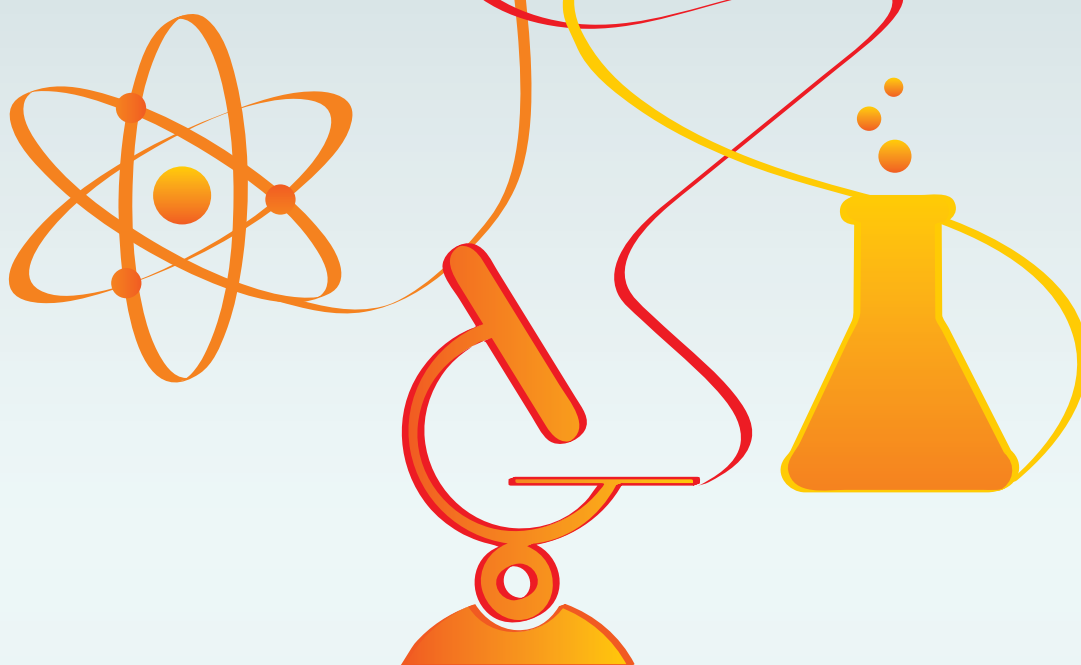
References

1. Grefsheim, S.F., Whitmore, S.C., Rapp, B.A., Rankin, J. A., & Robison, R. R. The informationist: building evidence for an emerging health profession, *J Med Libr Assoc*. 2010 April; 98(2): 147–156. [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859271/]

2. JISC, Virtual Research Environment programme, Web site accessed July 2010. [http://www.jisc.ac.uk/whatwedo/programmes/vre.aspx]

Note: The information in this article does not necessarily reflect the opinions of the National Institutes of Health. Any mention of a product or company name is for clarification and does not constitute an endorsement by NIH or the NIH Library. ❖