Volume 45 | Number 1 Article 10

7-2017

It Takes a Library to Preserve a Scientific Database: A Collaborative Exploration of Database Preservation

Bethany Anderson
University of Illinois at Urbana-Champaign, bgandrsn@illinois.edu

Tracy Popp
University of Illinois at Urbana-Champaign, tpopp2@illinois.edu

Electronic Currents

Assistant Editor: Joanne Kaczmarek, University of Illinois.
Contact Joanne at jkaczmar@illinois.edu if you would like to guest author a column or have a good idea to share.

It Takes a Library to Preserve a Scientific Database: A Collaborative Exploration of Database Preservation

By Bethany Anderson, University of Illinois, and Tracy Popp, University of Illinois

Scientists create a variety of digital assets as part of the research process. Some assets capture information about the world at a specific point in time; others attempt to test a hypothesis and reveal a "fact" about the world. The data generated by scientific endeavors, whether observational or experimental, manifest in a variety of formats and degrees of complexity depending on the discipline, the software, and the ways in which the data are intended to be used and reused by their creators and/or collaborators. The Archives at the University of Illinois at Urbana-Champaign (U of I) is currently working to preserve and make accessible one such digital asset, the International Registry of Reproductive Pathology database from the College of Veterinary Medicine. This database exemplifies the complexity of scientific data creation as well as underscores the importance of drawing on the expertise of archivists, digital preservationists, data curators, and subject specialists to address the challenges of preserving complex scientific digital objects.

Origins of the International Registry of Reproductive Pathology

The International Registry of Reproductive Pathology (the Registry) comprises a large hybrid scientific collection created by Kenneth B. McEntee, a veterinary pathologist who spent his career studying diseases of the reproductive system. Known for his meticulous recordkeeping, McEntee amassed data on about 20,000 cases of reproductive pathologies in animals.[1] McEntee collected most data during the 1970s before retiring from Cornell University where he had served as chair of the Department of Large Animal Medicine, Obstetrics and Surgery in the College of Veterinary Medicine. Shortly after retiring, McEntee moved himself and the Registry to the College of Veterinary Medicine (Vet Med) at U of I where he continued work on the Registry for another six years. After McEntee, pathology professor George Foley managed the Registry. During the late 1980s through the 1990s, the Registry drew reproductive pathologists from across the United States while also being used as a teaching collection. Given the number of veterinary pathologists and other scientists interested in accessing the Registry, Foley had hoped to make portions of the collection more accessible through digitization and Internet-based distribution, but his plan was never realized.[2] The Registry subsequently fell into disuse due to difficulties accessing and storing such a large, disparate collection of materials.

The Collection

The 20,000 cases in the Registry are documented by an array of materials, including wet tissue samples and sections, tissue slides, tissue samples encased in paraffin wax, typed case file reports, and a 1970s FoxPro database containing a searchable catalog of all case records. The database was indexed using the Systematized Nomenclature of Medicine (SNOMED), a standard for human and veterinary medicine terminology. McEntee and his collaborators noted that the Registry

> required indexing for ease of access to the details recorded in the files. For this purpose, terms depicting both general and specific concepts are required, as well as a multi-axial system for combining topography, morphology and etiology as needed. SNOMED meets these specifications more closely than any other system of which we are aware.[3]

Typed case files of the International Registry of Reproductive Pathology

Collaborative Approaches to Stewarding Scientific Records

The University of Illinois Archives has a long history of stewarding and preserving the records of science and technology, originating with the work of the first university archivist, Maynard Brichford.[4] Scientific and technological materials acquired by the archives include administrative records of departments and units, faculty papers, and scientific collections created by one creator but reused or added to over time by multiple collaborators and/or researchers. The archives has well-established procedures for acquiring and curating faculty papers proper, but scientific collections are often the products of collaborative scientific research having multiple creator(s)/user(s) over time and thus warrant different approaches to curation. With the recent launch of the University of Illinois Library's Research Data Service (RDS), stewarding and leveraging the U of I's digital (and often scientific) data sets has gained new visibility. The archives thus saw the Registry as a unique opportunity to leverage both the collection's digital data- and paper-based materials (i.e., typed case files) through a collaborative curatorial effort with the RDS. As a first step toward making the Registry accessible again, the FoxPro database was converted to Microsoft Access.

Following the appraisal of the paper-based and digital components of the collection by Bethany Anderson, William Maher, and Joanne Kaczmarek in October 2016, Anderson reached out to digital preservation coordinator Tracy Popp and the RDS staff to develop a plan for preserving the Registry. Anderson and Popp met with members of the RDS—Heidi Imker, Elizabeth Wickes, and Elise Dunham—and Susan Braxton, Prairie Research Institute librarian, who had recently attended a SIARD (Software Independent Archiving of Relational Databases) database preservation workshop. The team developed a plan to create access to digital object descriptions of the database through multiple access points, in both the archives' and RDS's respective repositories (which would be cross-linked to each other).[5] Additionally, they discussed methods and approaches to preserve and make accessible the database file. The team decided to explore whether the SIARD format could serve as a long-term preservation solution for the database.

Preservation Challenges

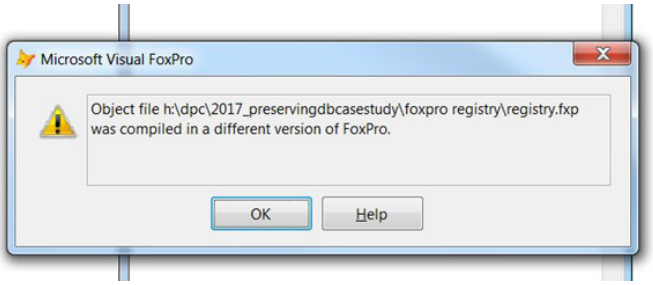
SIARD is meant to be format neutral, enabling curators to convert databases to a format that does not rely on proprietary software. SIARD tools enable conversions to and from a variety of database types (e.g., Oracle, MySQL, and Microsoft Access). A Databases Visualization Toolkit further makes it possible to query, explore, and export content from a database in the SIARD format.[6]

Using the SIARD Suite and Toolkit,[7] the team attempted first to convert the Access version of the database to the SIARD format. The conversion seemed to go fairly smoothly, but a few errors in the conversion log made the team pause. Additionally, during the appraisal of the collection, Vet Med personnel were not entirely confident that all data had been copied to Access from the FoxPro database. The team decided to access a copy of the original FoxPro database to see if any data loss had occurred.

Accessing the original FoxPro files turned out to be particularly challenging for several reasons. Not exclusively a database management environment, Visual FoxPro is also a dynamic programming language. Thus, anyone reviewing the files needs to understand which files contain application functionality and which files contain data to determine significant properties of the files individually and as a whole within the collection. One also needs access to the Visual FoxPro software. The U of I campus IT unit has an agreement with Microsoft which allows access to many software packages, though only those still supported. Thus, access was only available to the last version of Microsoft Visual FoxPro 9.0 (MSVFP) which was released in 2007. Unfortunately, MSVFP cannot access all of the files within the collection as they were created with an earlier version of the software. Specifically, MSVFP produced an error when a compiled program file was opened, indicating it could not be opened as it had been compiled in a different version of the software. Reviewing this compiled program file would likely offer clues as to how the program functioned and what may be important to migrate forward to retain functionality.

(Continued on page 28)

(Continued from page 27)



Error generated by Microsoft Visual FoxPro opening a compiled program file

The team's lack of knowledge compounded this challenge; to understand the various files in the collection and whether or not they are required to retain software functionality, a fair amount of research would need to be undertaken to understand basic MSVFP. For example, 19 DBX (or database) files exist within the collection. Five of the DBX files have the same name as five CDX files. Is there a relationship between these files and, if so, do the CDX files need to be retained for the data to be accessible and accurate if migrated to a contemporary software environment or file format?

Preliminary investigations indicate the database is a collection of unrelated tables. Any structure or relations seem to be established through two queries that have been migrated to SQL. Because the SIARD package relies on the SIARD software (i.e., we don't get any functionality from packaging the database in SIARD), we are preserving the database as flat file for access purposes. However, SIARD exports structural metadata to an XML file which is useful to ingest with the archival information package into our preservation repository.

Conclusion

Preserving complex digital scientific assets can certainly pose many challenges. This project has raised many questions about what researchers will need to understand the original databases, how to create access to them, and how to develop digital curation workflows that incorporate emerging best practices and formats afforded by the SIARD community as well as means for performing quality assurance checks. While the preservation of the database is still a work-in-progress, the possibilities of enhancing access to this resource wouldn't be within sight without the expertise and sharing of knowledge afforded by this interdisciplinary team of colleagues.

Notes

1. Howard E. Evans, Robert O. Gilbert, Bud C. Tennant, Donald H. Schlafer, "Kenneth B. McEntee," Cornell University Faculty Memorial Statement, Cornell University eCommons, hdl.handle.net/1813/18333.
2. Tania Banak, "Reproductive Pathology Samples Provide Valuable Information," *Veterinary Report* 19 (Fall 1995), IDEALS, hdl.handle.net/2142/89571.
3. D. O. Cordes, K. L. Limer, and K. McEntee, "Data Management for the International Registry of Reproductive Pathology Using SNOMED Coding and Computerization," *Veterinary Pathology* 18 (1981): 343.
4. Maynard Brichford, *Scientific and Technological Documentation: Archival Evaluation and Processing of University Records Relating to Science and Technology* (Urbana: University of Illinois, 1969), archives.library.illinois.edu/workpap/Sci-Tech-Documentation.pdf.
5. The University of Illinois Archives has recently transitioned to a new digital library platform for its born-digital and digitized collections, digital.library.illinois.edu; the RDS makes data sets accessible through the Illinois Data Bank, databank.illinois.edu.
6. Database Visualization Toolkit, visualization.database-preservation.com.
7. "SAIRD Suite," Swiss Federal Archives, www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html.