# The Art of Preserving Scientific Data: Building Collaboration into the Preservation of a Legacy Database

## Authors

Bethany Anderson, University of Illinois Archives, bgandrsn@illinois.edu*
Susan Braxton, Prairie Research Institute, braxton@illinois.edu*
Heidi Imker, Research Data Service, imker@illinois.edu*
Tracy Popp, Preservation Services, tpopp2@illinois.edu*

*University of Illinois at Urbana-Champaign, 1408 W. Gregory Drive, Urbana, IL 61801

## Abstract

Scientists collect, generate, and analyze data in a variety of forms. Oftentimes, this leads to datasets that have been organized, rendered, and stored by and in different methods, software, and storage locations. Databases are one means by which to organize and retrieve data. The complexity of databases as digital objects presents practical challenges for archivists, data curators, and digital preservationists. In this paper we discuss a collaborative project to assess and develop strategies for preserving and creating access for a scientific database from the veterinary medicine domain that leverages archival, digital preservation, and data curation expertise and resources.

**Keywords:** databases, scientific data, digital preservation, archives, appraisal, veterinary medicine, data sharing

## 1 Introduction

Scientists collect, generate, and analyze data in a variety of forms. Oftentimes, this leads to datasets that have been organized, rendered, and stored by and in different methods, software, and storage locations. Databases are one means by which to organize and retrieve data. Given increased recognition that data are critical research products worthy of curation, preservation, and long-term stewardship, the complexity of databases as digital objects presents practical challenges for archivists, data curators, and digital preservationists.

Between 2016 and 2017, the University of Illinois Archives acquired the hybrid research records of the International Registry of Reproductive Pathology project. The collection includes data collected by veterinary pathologist Dr. Kenneth B. McEntee for 20,000 cases of reproductive pathologies in mammals. In addition to typed case files that contain correspondence, reports, and photographs, the collection includes physical specimens and a FoxPro database containing a catalog of all case records. The registry database dates to the late 1970s and early 1980s as the pathology community worked to harness the power of computers and database technologies to facilitate data organization and retrieval [3]. Indexed using the Systematized Nomenclature of Medicine (SNOMED),[1] the database's importance is twofold: as a historical example of technology and standards adoption in the veterinary medicine community but also for the value of the data held therein.

Because the database was acquired in the FoxPro format, it could not be easily accessed or appraised for enduring archival value without first assessing its structure and the native software used to organize and store the data in this digital object. The database raised questions about the object and its relationship to its native software and what users would need to access, understand, and reuse the data, as well as how to appraise and develop digital curation workflows for databases. Creating a strategy for preservation and access for the database required the expertise of multiple people within the University of Illinois Library. In this paper, we discuss a collaborative and interdisciplinary project to assess and develop strategies for preserving and creating access for digital scientific assets that leverages archival, digital preservation, and data curation expertise and resources.

[1] See SNOMED: https://www.snomed.org/.

## 2 Collection History and Overview

Over several decades, veterinary pathologist Kenneth B. McEntee collected data on approximately 20,000 reproductive pathologies in animals. The result was the International Registry of Reproductive Pathology (called "registry" hereafter), a hybrid scientific collection consisting of paper-based typed case files, physical specimens, and a born-digital database. McEntee was a specialist in reproductive pathologies in mammals and spent the majority of his career at Cornell University (1947-1980), where he served as professor and chair of the College of Veterinary Medicine's Department of Large Animal Medicine, Obstetrics and Surgery [4]. McEntee retired from Cornell in 1980 but continued to work on the registry for another six years as a visiting professor of reproductive pathology at the College of Veterinary Medicine (Vet Med) at the University of Illinois at Urbana-Champaign (U of I). The collection was used for teaching and research, though these uses decreased over time (Vet Med staff, pers comm., October 21, 2016). Vet Med maintained the collection until the transfer of the typed case files and database to the University of Illinois Archives in 2016 and 2017.[2]

McEntee was known for his "painstaking clinical examination coupled with detailed gross and microscopic examination, meticulous record keeping and unceasing deliberation" [4]. The registry's 20,000 cases are documented by materials in a number of different formats, including wet tissue samples and sections, tissue slides, tissue samples encased in paraffin wax, typed case file reports, and the FoxPro database containing a searchable catalog of all case records. The database was indexed using SNOMED, a standard for human and veterinary medicine terminology [3]. Data created with the care and diligence as that done by McEntee are rare, which increases the intrinsic value of the collection.

## 3 Developing a Preservation Strategy

The Archives reached out to the Preservation Services unit and the Research Data Service (RDS) to review the database and develop a strategy for its preservation. The Archives acquired the relational database in two formats: its native FoxPro format and Microsoft Access (its tables had also been migrated to by Vet Med to facilitate use). Archives, Preservation, and RDS staff enlisted the expertise of a library colleague who had recently attended a "Software Independent Archiving of Relational Databases" (SIARD) preservation workshop at the 2016 iPRES conference.[3] SIARD is an open preservation format for relational databases created by the Swiss Federal Archives. The team of four, representing the authors of this publication, agreed that the database could be made accessible through multiple access points in both the Archives' and RDS' content management systems, while also serving as a precedent for future historical datasets that could be jointly stewarded by both the Archives and RDS. We decided to explore whether the SIARD format could serve as a long-term preservation solution for the database and create two access points for the database through the Archives' and the RDS' systems.

## 4 Database Preservation

Relational databases have been used widely across many research disciplines for decades, with software products available expanding considerably in the early-1990s [6, 7], and their preservation is a known challenge. Recommended practices exist for archiving records from actively used databases [10], but archivists are concerned with preservation of the structure and content databases, not just the data. Archivists and digital preservationists have reported case studies of emulation [2] and export/migration to either commercial or open database formats [9]. These cases aim to preserve the database as is—to preserve the structure, metadata, and at least some of the functionality of the database. Somewhat in contrast, research data management and preservation recommendations may encourage researchers to migrate data to preservation-friendly formats [1]—a common example being CSV for tabular data—for long-term preservation, deposit into a repository, and/or to maximize long-term reusability by others.

[2] The University of Illinois Archives only acquired the paper-based and born-digital components of the collection; the rest of the materials were kept by Vet Med.

[3] Luis Faria, Marcel Büchler, and Kuldar Aas, "Relational Database Preservation Standards and Tools," October 6, 2016, iPRES 2016, http://www.ipres2016.ch/frontend/index.php?page_id=2833.

## 4.1 SIARD Format Exploration

Using instructions and software provided at the iPRES 2016 workshop on preservation of relational databases [5], the team met to explore conversion of the registry database to SIARD 2 format, motivated primarily by the possibility of querying the database within the University of Illinois Library's content management systems. Since SIARD does not support the FoxPro format, the Microsoft Access version was converted to SIARD 2 using the SIARD Suite. We found the conversion log provided a useful summary of the registry's twenty tables and provided a deeper understanding of the database structure than the team had been able to glean from inspecting the database in Access. The team opted to not retain the SIARD file format for preservation due to the lack of relations among data tables and difficulty opening the SIARD file created during testing; metadata SIARD produced was retained and included in the preservation description information in the Submission Information Package (SIP) as it contained information about table names, fields, and record counts—metadata which may prove useful to future users. Thus, SIARD proved to be a useful appraisal tool, but it did not provide any functionality that would enhance access for users.

## 5 Digital Preservation Challenges

Following preliminary testing with the SIARD Suite, the Library's digital preservation coordinator (DPC) used the SIARD conversion log and Microsoft Access version for reference and based preservation decisions on the FoxPro version. We were uncertain whether all data had been successfully migrated from FoxPro to the Microsoft Access version, and so the team decided to focus on migrating data from the original FoxPro format. Rendering the native FoxPro files proved challenging. Although the DPC has generalized technical knowledge of relational database structure, creation, and development principles, gaining knowledge to understand FoxPro operation and idiosyncrasies to determine the files' significant properties and dependencies required time-intensive research about the software. Members of the U of I Library's IT staff were consulted for additional technical expertise and software procurement.

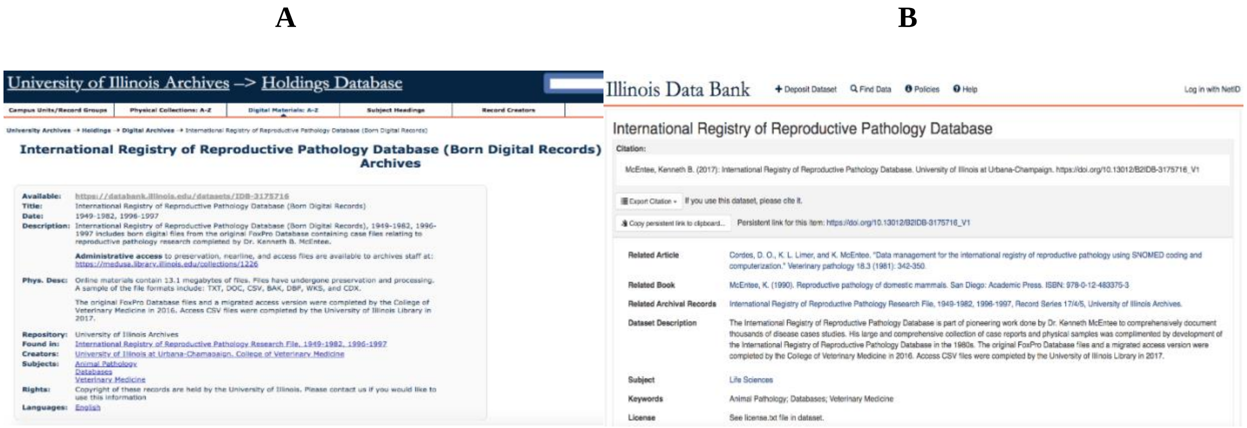
The best option for rendering the files was the last version of Microsoft Visual FoxPro 9.0 (MSVFP) which was released in 2007. A U of I campus-negotiated agreement provides use of many Microsoft applications; however, access is limited to actively supported programs. MSVFP was the only FoxPro version still supported by Microsoft. An error message when attempting to open one of the FoxPro compiled program files provided a clue that MSVFP was not the version used to create this iteration of the database.

## 5.1 Preservation and Access Decisions

After further investigating FoxPro documentation, reviewing the Microsoft Access database, and concurrently investigating output from the SIARD application, the DPC realized that pertinent data was stored in DBF files (table data) which were renderable in MSVFP. Upon further review, she discovered that relations between the tables were established through queries—the database did not contain explicit relations between tables. Following the Library of Congress' "Recommended Format Statement for Datasets/Databases" [8], the team decided to make the database accessible as discrete CSV flat files exported from the DBF files.

Given the Archives and the RDS' different areas of focus, each unit offers resources with strengths that align with their unique missions and expected user communities. We sought to capitalize on those strengths and ingested the database files into both the Archives' collection in the digital preservation repository, Medusa, and the Illinois Data Bank. The Medusa repository platform, which manages and tracks master files across three distinct and geographically distributed storage locations and serves access copies, is also the base technology for the Illinois Data Bank. The Archives took initiative to prepare, describe, and ingest all files into Medusa. The resulting SIP that was ingested comprised the original DBF files and Microsoft Access version as the preservation masters, the CSV files as the access copy, and the metadata exported from SIARD (and other contextual information about the conversion process) as part of the preservation description information. The Archives represents the database as part of the larger hybrid collection including paper-based and digital components, offers mediated access to the preservation, nearline, and access files, and links to the corresponding dataset in the Illinois Data Bank (Figure 1A). Likewise, the team member from the RDS took initiative to prepare a deposit for the Illinois Data Bank. Documentation in the form of a ReadMe file was created which contains additional information; for example, current scientific contacts within the College of

Veterinary Medicine. This documentation was deposited into the Illinois Data Bank along with the XML metadata and CSV files. The Illinois Data Bank provides a DOI with additional discoverability through registration in DataCite, offers unmediated access to the CSV and documentation files, and likewise directs users to the full collection held by the Archives (Figure 1B).[4] Thus, the two records were created with the intent to complement and refer to each other. As each individual from Archives and RDS worked on ingest, they communicated with each other to harmonize the records. For example, licensing practices are especially different between the two units. Given the broad range of materials found within the Archives, "all rights reserved" is common. However, emerging data publication standards encourage licenses to be as permissive as possible and many data repositories prefer or require CC0. Applying distinct licenses would have been potentially problematic, and this provided an especially practical learning opportunity between the units. Given the nature of the larger collection within the Archives, the more restrictive license was selected.


A
B
University of Illinois Archives –> Holdings Database
Campus Units/Record Groups
Physical Collections: A-Z
Digital Materials: A-Z
Subject Headings
Record Creators
University Archives → Holdings → Digital Archives → International Registry of Reproductive Pathology Database (Born Digital Records)
Illinois Data Bank
Deposit Dataset
Find Data
Policies
Help
Log in with NetID
Export Citation
Copy persistent link to clipboard...

International Registry of Reproductive Pathology Database (Born Digital Records) Archives

| | |
|---|---|
| Available: | https://databank.illinois.edu/datasets/IDB-3175716 |
| Title: | International Registry of Reproductive Pathology Database (Born Digital Records) |
| Date: | 1949-1982, 1996-1997 |
| Description: | International Registry of Reproductive Pathology Database (Born Digital Records), 1949-1982, 1996-1997 includes born digital files from the original FoxPro Database containing case files relating to reproductive pathology research completed by Dr. Kenneth B. McEntee.<br>**Administrative access** to preservation, nearline, and access files are available to archives staff at: https://medusa.library.illinois.edu/collections/1226 |
| Phys. Desc: | Online materials contain 13.1 megabytes of files. Files have undergone preservation and processing. A sample of the file formats include: TXT, DOC, CSV, BAK, DBF, WKS, and CDX.<br>The original FoxPro Database files and a migrated access version were completed by the College of Veterinary Medicine in 2016. Access CSV files were completed by the University of Illinois Library in 2017. |
| Repository: | University of Illinois Archives |
| Found in: | International Registry of Reproductive Pathology Research File, 1949-1982, 1996-1997 |
| Creators: | University of Illinois at Urbana-Champaign. College of Veterinary Medicine |
| Subjects: | Animal Pathology<br>Databases<br>Veterinary Medicine |
| Rights: | Copyright of these records are held by the University of Illinois. Please contact us if you would like to use this information |
| Languages: | English |

International Registry of Reproductive Pathology Database

Citation:

McEntee, Kenneth B. (2017): International Registry of Reproductive Pathology Database. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-3175716_V1

If you use this dataset, please cite it.

Persistent link for this item: https://doi.org/10.13012/B2IDB-3175716_V1

| | |
|---|---|
| Related Article | Cordes, D. O., K. L. Limer, and K. McEntee. "Data management for the international registry of reproductive pathology using SNOMED coding and computerization." Veterinary pathology 18.3 (1981): 342-350. |
| Related Book | McEntee, K. (1990). Reproductive pathology of domestic mammals. San Diego: Academic Press. ISBN: 978-0-12-483375-3 |
| Related Archival Records | International Registry of Reproductive Pathology Research File, 1949-1982, 1996-1997, Record Series 17/4/5, University of Illinois Archives. |
| Dataset Description | The International Registry of Reproductive Pathology Database is part of pioneering work done by Dr. Kenneth McEntee to comprehensively document thousands of disease cases studies. His large and comprehensive collection of case reports and physical samples was complimented by development of the International Registry of Reproductive Pathology Database in the 1980s. The original FoxPro Database files and a migrated access version were completed by the College of Veterinary Medicine in 2016. Access CSV files were completed by the University of Illinois Library in 2017. |
| Subject | Life Sciences |
| Keywords | Animal Pathology; Databases; Veterinary Medicine |
| License | See license.txt file in dataset. |

Figure 1. Panel A shows a screenshot of the record as represented in the Archives collection. Panel B shows a screenshot of the record as represented in the Illinois Data Bank.

## 6 Collaborative Preservation

By virtue of the productive working relationships fostered by the collaborative nature of the U of I Library, it was natural for the team to connect with each other to tackle preservation of the database. Ultimately, the individuals included in the collaborative team included an expert in archives, an expert in digital preservation, a library colleague with both database experience and SIARD exposure, and an expert in research data publishing. While each individual has various levels of familiarity with the other units, no one person was sufficiently deeply versed in each area to move the project to full completion. Although theoretically such deep knowledge could be found in a single individual, we believe these persons are generally referred to as "unicorns." Indeed, one concern that emerged from the team in hindsight was if all the necessary expertise had been consulted. For example, one could also imagine a role for the liaison librarian for Vet Med. This highlights the collaborative challenge of balancing the need to include all relevant parties without sacrificing the ability to make progress on a reasonable timescale.

Regardless of these challenges, it became immediately clear that by virtue of our respective roles within the Library, we approached the project in different, but complementary, ways. For the colleague with SIARD exposure, a pressing question was how does this database actually function—i.e., what are the relationships between the tables? For the archivist, ensuring the database would be represented as one component in an otherwise complex collection with physical samples, paper files, etc., was paramount. The digital preservation

[4] The database is described and its files are accessible in the RDS's Illinois Data Bank, https://doi.org/10.13012/B2IDB-3175716_V1, and described in the University of Illinois Archives' content management system, https://archives.library.illinois.edu/archon/index.php?p=digitallibrary/digitalcontent&id=10813. Both descriptive records cross-link to each other.

coordinator was consumed with determining how we could open, understand, and reformat these files to enable a more durable future. Finally, as a position steeped in Open Data, the staff from the Research Data Service was instinctively concerned that the resulting data would adhere to the FAIR principles for scientific data to ensure maximum dissemination and downstream utility [11]. It's unlikely an "ideal" team to address such challenges will ever exist, in theoretical or practical terms. However, through exposure to each other's questions and perspectives, we gained the ability to be more cognizant of a multifaceted approach and then be able to further assess whose expertise is needed (and available).

## 7 Conclusion

The team assembled to address the preservation of this database represented multiple areas from the University of Illinois Library. Ultimately, this project not only resulted in the preservation of an important data collection, which proved to be a highly illuminative exercise in and of itself, but it also enabled a deeper understanding of each other's work *in practice*. By following each other through the process and working either in tandem or sequentially through activities at different stages, we were able to shadow each other and learn more deeply about each other's expertise and our respective unit's operations, norms, and values. This provided a unique opportunity to understand how a collaborative approach to preservation can function in reality. Our initial work thus helped us understand the expertise needed and the overall distribution of effort required, which was not necessarily even in this case, such that our future efforts to work collaboratively to preserve valuable data collections will continually improve.

## Acknowledgments

The authors would like to thank the U of I College of Veterinary Medicine for providing information about the history of the registry and collaborating on the preservation of the database.

## References

[1] Kristin Briney. 2015. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. Exeter, UK: Pelagic Publishing.

[2] Euan Cochrane, Dirk von Suchodoletz, and Mike Crouch. 2013. Database Preservation Using Emulation–a Case Study. *Archifacts*, 80–95.

[3] D. Cordes, K. Limer, and K. Mcentee. 1981. Data Management for the International Registry of Reproductive Pathology Using Snomed Coding and Computerization. *Veterinary Pathology* 18, no. 3, 342–50. DOI: https://doi.org/10.1177/030098588101800307.

[4] Howard E. Evans, Robert O. Gilbert, Bud C. Tennant, Donlad H. Schlafer. 2005. Kenneth B. McEntee. Cornell University Faculty Memorial Statement. http://hdl.handle.net/1813/18333.

[5] Luis Faria, Marcel Buchler, and Kuldar Aas. 2016. Relational Database Preservation Standards and Tools workshop. *IPRES 2016 - 13th International Conference on Digital Preservation*. Berne, Switzerland. http://www.ipres2016.ch/frontend/index.php?page_id=2833.

[6] Thom Gillespie. 1991. Databases: In Transition (book review). *Library Journal* 116, no. 21, 188.

[7] Clifton Karnes. 1993. Editorial License. *Compute!*. https://www.atarimagazines.com/compute/issue138/4_Editorial_license.php.

[8] Library of Congress. Recommended Formats Statement. Accessed April 13, 2018. https://www.loc.gov/preservation/resources/rfs/data.html.

[9] Andrew Lindley. 2013. Database Preservation Evaluation Report-SIARD vs. CHRONOS. *10th International Conference on Preservation of Digital Objects* 29. DOI: 10.13140/2.1.3272.8005.

[10] Jack E. Olson. 2008. *Database Archiving How to Keep Lots of Data for a Very Long Time*. Burlington, MA: Morgan Kaufmann/Elsevier.

[11] Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 1-9. DOI: http://dx.doi.org/10.1038/sdata.2016.18.