ISSN 2161-3974

JeSLIB 2018; 7(3): e1154
doi:10.7191/jeslib.2018.1154

Journal of eScience Librarianship
putting the pieces together: theory and practice

EScience in Action

Behind eMammal's Success: A Data Curator With A Data Standard

Jennifer Y. Zhao and William J. McShea

Smithsonian Institution, Front Royal, VA, USA

Abstract

This paper explores the data challenges of a major collection method in the field of ecology: using infrared-activated cameras to detect wildlife. One such solution, eMammal, is now available to address these struggles. We delineate the key reason behind its success: a data curator who manages an established data standard and communicates with eMammal's users and stakeholders. We outline the tasks of this data curator, mention how they can work with data librarians, and demonstrate that the data curator position is already applicable in several biological science fields with a few examples. We end by emphasizing the growth of such a position and how it contributes to the research field.

Correspondence: Jennifer Y. Zhao: ZhaoJJ@si.edu
Keywords: camera trap, data curator, data management, data standard, ecology

Disclosures: The substance of this article is based upon a lightning talk presentation at RDAP Summit 2018. Additional information at end of article.

JeSLIB 2018; 7(3): e1154
doi:10.7191/jeslib.2018.1154

Background

Monitoring the world's biodiversity has traditionally been conducted through the collection and curation of physical specimens in museums. Here, we define curate as organizing and overseeing for additional usages, such as public exhibition or use for data analysis (*American Heritage Dictionary of the English Language*). Combined with written tags containing detailed information incorporating geolocations, date, et cetera, these items are valuable historical records that document changes in species' occurrence and distribution (McCarthy 1998, 9; Shaffer et al. 1998, 27). However, modern day physical specimens become more and more difficult to obtain due to permitting issues, expense, and lack of trained personnel. In addition, the public has called for noninvasive sampling techniques due to animal welfare concerns (O'Connell et al. 2011, 3).

New technology, called camera traps, has emerged which can resolve many of these traditional collection issues. Current camera traps are infrared-activated sensors connected to digital cameras that can be placed in the wild as a non-invasive method of monitoring many mammal species (Kays and Slauson 2008, 110-111). When the wildlife image is tagged with required metadata (location, date, timestamp and species ID), it becomes the near equivalent of the traditional museum specimen (Rovero and Zimmermann 2016, 3; Fegraus et al. 2011, 345). It is not an exact equivalent because with camera traps, the sampling does not include key physical properties such as genetic material. Despite this, scientists are increasingly turning to camera traps as their main collection method.

However, after switching to collecting camera trap data, scientists realized this shift from traditional to digital specimens was not so simple. Digital collection involved an entirely new suite of logistical limitations that had to be overcome to make camera traps a truly effective tool for monitoring the global distributions of mammal species. One example of a logistical difficulty was the collection rate of specimens per year evolved from hundreds of physical specimens to tens of thousands of digital specimens, making the datasets challenging to manage (Meek et al. 2014, xi).

This is a struggle familiar to many in the world of eScience and research data (Lord et al. 2004, 371). For this particular problem, eMammal (https://emammal.si.edu) was designed to address the challenges in managing substantial collections of digital wildlife specimens.

What is eMammal?

eMammal is an end-to-end system for gathering, curating, and sharing camera trap images and data. This platform was created to address the camera trap community's need for a data management solution. William McShea, Robert Costello, and Roland Kays collaborated on a pipeline of software for processing, storing, and analyzing wildlife data based on images (McShea et al. 2015, 57-60). The first challenge was to ensure photos were tagged with required metadata—a species identification with a location, a count of the number of this species present, and a date and time stamp. The second challenge was to create a data standard and schema for consistency in the data collection of researchers and projects (Forrester et al. 2016, e10197). With Excel spreadsheets, for example, there is no validation that data integrity and consistency is kept throughout the data processing stage for any camera trapping project with significant amounts of data. eMammal, created in conjunction with

JeSLIB 2018; 7(3): e1154
doi:10.7191/jeslib.2018.1154

Smithsonian researchers, Smithsonian staff from the Research Computing Department, and researchers from the North Carolina Museum of Natural Sciences, is the answer to these challenges. The eMammal data pipeline consists of four components: a desktop application for the initial tagging and uploading of photos, a cloud-based expert review tool for verifying the photo tags, a repository to archive and store these images, and a website that serves as both a project management tool and a data serving portal (see Figure 1). In this data pipeline, the data standard (outlined further in Forrester et al. 2016) is presented through an easy and intuitive interface. The streamlining of the process makes data management the logical solution for large camera trapping projects. Currently, eMammal supports 110 different research projects, with over 70 projects currently collecting new specimens, and over 860,000 animal detections from 22 countries (Smithsonian Institution 2018).

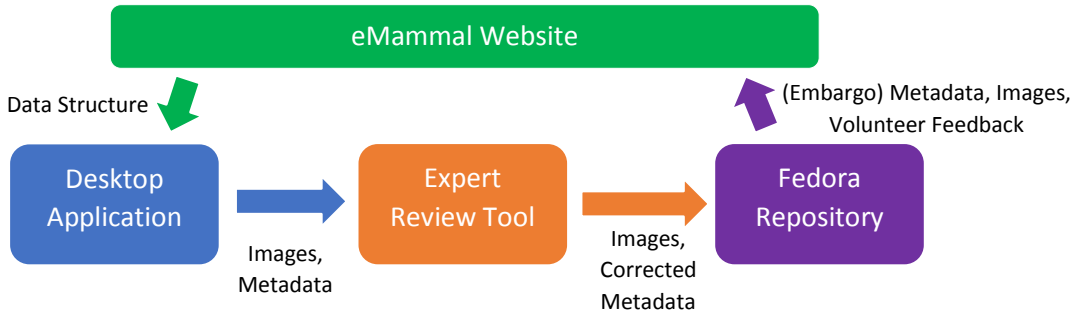

Figure 1: Simplified schematic of the eMammal pipeline. The eMammal Website provides the basic data structure, set up by each project manager, for the Desktop Application. Staff and volunteers then tag sets of images with metadata on the Desktop Application, which flow to the Expert Review Tool (ERT). On the ERT, experts verify image tags, animal counts, camera setup, and camera geographical placement. All data then flow to the Fedora-based Data Repository, where images and metadata are backed up according to best practices for research data management. Finally, the eMammal Website extracts data from the repository for staff and volunteer feedback and for the general public, with the exception of sensitive data such as endangered species location and temporary embargoes imposed for researchers to publish their findings.

Why is eMammal successful?

eMammal serves as an example of pioneering good data standards through early establishment and enforcement of data compliance. This required extensive meetings with both camera trap researchers and information technology-based research data experts. Careful planning in software design was required to meet the minimum requirements of all possible individual projects. Enforcing full compliance with these data entry standards through validation then ensures the resulting dataset is globally consistent, with minimal "variability and errors" (Meek and Zimmermann 2016, 221).

This process leads to the need for a data curator who understands the data standard and pipeline. For eMammal, the data curator's role is to not only organize and groom the eMammal dataset, but to also support the pipeline and its many users. The duties of a data curator manifest in multiple ways and are expanded upon below.

JeSLIB 2018; 7(3): e1154
doi:10.7191/jeslib.2018.1154

The core task of a data curator is to ensure that the data standard is upheld. A common issue with the data standard is a researcher requesting additional fields for their eMammal project. To uphold the data standard, the data curator must work with all parties involved with the project in question. The most common resolution is that one of the variables in the existing data standard nullifies the need for an additional field. Less frequently, the data curator may find that this request for a particular variable would benefit many other research projects and ultimately work to incorporate it as a new field into the data standard.

As a second core task, the eMammal data curator supports the eMammal pipeline by maintaining open communication between all eMammal stakeholders, including but not limited to: researchers, research volunteers, information technology personnel, and software developers. On the research side the data curator coordinates researchers and their volunteers—termed citizen scientists—by handling any software issues by communicating its limitations and addressing any misconceptions about its capabilities. In turn, the researchers suggest improvements to the data curator. To then effectively pass along these communicated needs, the data curator serves as a translator for information technology personnel and software developers. Researchers and computer programmers differ in their fundamental vocabulary, and a data curator must be familiar with both vocabularies so as to increase the ease and openness of communication (Leonelli 2016, 33). For example, the computer programmers were instructed to create bar graphs on "detection rate." To ecologists, the word "rate" is interpreted as some value per unit of measure (e.g. time or space), so naturally the researchers assumed that the programmers would understand to divide by time. However, the programmers returned with the raw total number of detections. The data curator caught this mistake in the early stages of programming, and then requested clarification from the researchers regarding what unit of time to divide by, and to provide a sound definition of "detection rate." If the programmers did not understand the term, than the general public audience may not as well, and so a definition on the public-facing website would be helpful to all.

The data curator also speaks with the host institution's supporting departments for information technology matters, such as the Office of the Chief Information Officer (OCIO) and Smithsonian Libraries. The data curator collaborates with OCIO to ensure the integrity of the backup copies of all eMammal research data. They can help address any problems that surface and additionally report new problems that OCIO may overlook in their management of the data repository. On the other hand, the connection with the Smithsonian Libraries was vital in implementing the creation of DOIs for eMammal researchers. The data curator is also able to contribute to the Smithsonian Libraries' Data Management Working Group by providing real world examples of implementation and researcher wrangling, such as helpful ways of framing the need for data management.

A third task of eMammal's data curator is to oversee data quality. With such a large dataset, it takes both the researcher sighting inconsistencies in their data and the data curator's spot checks to ensure data quality. Data quality is also affected by improvements and maintenance of the code behind the pipeline. The data curator is required to continually seek pipeline improvements by staying abreast of new developments and scaling up the capabilities of the pipeline as needed. Along with improvements, software maintenance is another responsibility of the data curator, who has to implement any applicable updates in technology and security. However, as both maintenance patches and code improvements are applied, unexpected

inconsistencies may appear in the data. The data curator is expected to catch and fix these cases of data corruption once reported, and also set up contingency measures so as to prevent such incidents in the future.

Finally, eMammal's data curator expands the reach of the camera trap dataset by recruiting and training new project managers, and also seeking out legacy datasets to build out the data collection of the past. In this way, more researchers using camera trap technology will gain practice in data management. Recruitment of legacy data collections ensures a well-rounded dataset, and opens opportunities for larger, longer-term studies.

In addition to recruitment of legacy data collections, data curators must seek help from data librarians in recruiting current data collections. This complements the role of a data librarian well, as they are usually consulted on selecting appropriate data management tools. Data librarians can help the eMammal data curator by spreading the word about this technology and by recommending exploration of its use. Once a data librarian knows one of their researchers is planning to use, is currently using, or has finished using camera traps in their research they can recommend eMammal.

Conclusion

The utility of a data curator is not limited to eMammal, but is rather a more universal need. All across the biological sciences, researchers are transitioning to deal with big data and its challenges. Other prominent examples of curation needs are represented in similar data repositories, including Movebank which stores radiotracking locations (Kranstauber et al. 2011, 834) and GenBank which stores publicly available nucleotide sequences (Benson et al. 2009, D26).

This transition from physical, tangible curation to digital curation calls for a person whose role is to curate data but also incorporate ever-changing technology with limited resources. No longer can the curator rely on manual techniques of data processing, but must also understand technological trends and communicate them to the program users. The rate of this new technology development is becoming ever shorter as digital innovation leads to faster cycles of creative destruction, meaning data processing technology today may be obsolete tomorrow (Govindarajan and Srivastava 2016, 24-25). The new era of curators are responsible for keeping up with and incorporating the new technology. The curators must also work in tandem with data librarians, who are essential in outreach and recruitment of the usage of the new technology.

The job of a data curator is an increasingly essential role, as the position provides unique support for researchers and computer programmers. It takes the burden of keeping up with the latest data processing technology off of the researcher's shoulders, yet also lessens the burden of the computer programmers by condensing all requests for repair to improve the existing technology. The programmer is then able to pinpoint the origin of technological issues more efficiently and provide a fix while understanding what must be done to improve the system. Thus, the researcher is able to concentrate on furthering their research, and the community as a whole can continue its diverse research initiatives. Because eMammal has a data curator, many of these needs are met for the camera trapping community. Beyond this community of researchers, the data curator also keeps track of and informs efforts in the

creation of better data management platforms. A data curator has a foot in both worlds of research data and computer science, thereby connecting both worlds. This vital connection leads to a better way of addressing the incorporation of big data into the biological sciences and provides a direction for future research.

Acknowledgments

The authors thank Joseph Kolowski for lending some key reading material, and Sultana Majid, Kerith Wang, and Joyce Huang for their valuable feedback on a draft of this article. The authors also thank the reviewers for their valuable suggestions, as well as Justin Cooper. eMammal has various sources of funding, including NSF grants and gifts from donors and thanks them.

Disclosures

The substance of this article is based upon a lightning talk presentation at RDAP Summit 2018: "eMammal: Data Management Pipeline for Camera Traps" https://osf.io/skrz3.

References

American Heritage Dictionary of the English Language. 5th ed. "Curate." Boston: Houghton Mifflin Harcourt, 2011. Accessed September 10, 2018. https://www.ahdictionary.com/word/search.html?q=curate

Benson, Dennis A., Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2009. "GenBank." *Nucleic Acids Research* 37(1): D26-D31. http://doi.org/10.1093/nar/gkn723

Fegraus, Eric, Kai Lin, Jorge Ahumada, Chaitan Baru, Sandeep Chandra, and Choonhan Youn. 2011. "Data acquisition and management software for camera trap data: a case study from the TEAM Network." *Ecological Informatics* 6(6): 345-353. http://doi.org/10.1016/j.ecoinf.2011.06.003

Forrester, Tavis, Timothy O'Brien, Eric Fegraus, Patrick Jansen, Jonathan Palmer, Roland Kays, Jorge Ahumada, Beth Stern, and William McShea. 2016. "An Open Standard for Camera Trap Data." *Biodiversity Journal* 4: e10197. http://doi.org/10.3897/BDJ.4.e10197

Govindarajan, Vijay and Anup Srivastava. 2016. "The Scary Truth About Corporate Survival." *Harvard Business Review* December 2016: 24-25. https://hbr.org/2016/12/the-scary-truth-about-corporate-survival

Kays, Roland W. and Keith M. Slauson. 2008. "Remote cameras." In *Noninvasive survey for carnivores*, edited by Robert A. Long, Paula MacKay, Justina C. Ray, and William J. Zielinski, 110-140. Washington, DC: Island Press.

Kranstauber, Bart, Alison Cameron, Rolf Weinzerl, Tony Fountain, Sameer Tilak, Martin Wikelski, and Roland Kays. 2011. "The Movebank data model for animal tracking." *Environmental Modelling & Software* 26(6): 834-835. http://doi.org/10.1016/j.envsoft.2010.12.005

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago: The University of Chicago Press.

Lord, Phillip, Alison Macdonald, Liz Lyon, and David Giaretta. 2004. "From Data Deluge to Data Curation." In *Proceeding of the 3th UK e-Science All Hands Meeting* 371-375. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.111.7425&rep=rep1&type=pdf

McCarthy, Michael A. 1998. "Identifying declining and threatened species with museum data." *Biological Conservation* 83(1): 9-17. https://doi.org/10.1016/S0006-3207(97)00048-7

McShea, William J., Tavis Forrester, Robert Costello, Zhihai He, and Roland Kays. 2015. "Volunteer-run cameras as distributed sensors for macrosystem mammal research." *Landscape Ecology* 31(1): 55-66. http://dx.doi.org/10.1007/s10980-015-0262-9

Meek, Paul D. and Fridolin Zimmermann. 2016. "Camera traps and public engagement." In *Camera Trapping for Wildlife Research* 219-236. Exeter, UK: Pelagic Publishing.

Meek, Paul D., Peter J. S. Fleming, Guy Ballard, Peter Banks, Andrew W. Claridge, Jim Sanderson, and Don Swann. 2014. *Camera trapping: wildlife management and research*. Melbourne, Australia: CISRO Publishing.

O'Connell, Allan F., James D. Nichols, and K. Ullas Karanth. 2011. *Camera Traps in Animal Ecology: Methods and Analyses*. New York: Springer. http://doi.org/10.1007/978-4-431-99495-4

Rovero, Francesco, and Fridolin Zimmermann. 2016. *Camera Trapping for Wildlife Research*. Exeter, UK: Pelagic Publishing.

Shaffer, Howard B., Robert N. Fisher, and Carlos Davidson. 1998. "The role of natural history collections in documenting species declines." *Trends in Ecology and Evolution* 13(1): 27-30. https://doi.org/10.1016/S0169-5347(97)01177-4

Smithsonian Institution. 2018. "Map of Projects." Accessed September 11, 2018. https://emammal.si.edu/projects