DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2016

Poster

Modeling Cultural Evolution with Metadata Collections: A Phylomemetic Approach

Nicholas M. Weber
University of Washington, USA
nmweber@uw.edu

Andrea K. Thomer
University of Illinois at Urbana-Champaign, USA
thomer2@illinois.edu

Keywords: Evolution of metadata; collections; formal modeling; phylomemetics

1. Introduction

Descriptive metadata is typically used to record information about digital artifacts and thereby facilitate users' retrieval and use of these artifacts. The resulting collections of descriptive metadata records may be considered digital artifacts in and of themselves, and evidence of the behavior and values of the communities and cultures that produce, use, and cooperate in provisioning digital artifacts. Studying the ways that collections of metadata records change over time may reveal novel insights into the evolution of the communities that not only create digital artifacts – but that catalog and manage them as well.

In this poster we describe and apply an approach to modeling cultural evolution, *phylomemetic analysis,* using collections of metadata records. We show that collections of descriptive metadata records can be used as a primary data source for the evolutionary analysis of institutions and communities engaged in digital scholarship, and discuss the potential implications of this method for metadata repository managers and researchers alike.

2. Phylomemetics

Derived from (and named after) phylogenetic methods in evolutionary biology, phylomemetics refers to the evolutionary analysis of non-genetic or biological data (Howe & Windram, 2011). In a phylogenetic analysis of biological specimens unique aspects of an organism (e.g. DNA sequences, the number of toes on a limb; the presence or absence of a hair or feathers; or as Darwin himself demonstrated, the different shapes of birds' beaks) are coded qualitatively as *characters* and then statistically analyzed to infer an evolutionary tree. In a phylomemetic analysis, "memes" rather than genes are coded and analyzed. This approach has previously been used to study cultural evolution through a range of artifacts, both physical and conceptual (e.g. cornets (Tëmkin & Eldredge, 2007), arrowheads (O'Brien, Darwent & Lyman 2001), languages (Bates and Elman, 2000), music (Le Bomin, Lecointre & Heyer, 2016), and folk tales (Tehrani, 2013)).

Just as descriptive metadata from digital libraries, such as the HathiTrust, can be studied through "distant readings" of cultural trends over time (Underwood, 2016), so can collections of metadata records. For example, in previous work we've shown that phylomemetic methods can be applied to collections of NASA metadata records by using attribute-value pairs as characters; in doing so, we are able to identify clusters of user communities altering an earth science dataset for similar purposes, as well as points at which communities split apart from one another (Thomer and Weber, 2014). Thus, descriptive metadata collections can be used to model cultural change within communities that produce, share, and alter datasets and other digital artifacts. Though this change is often self-reported to a degree through texts such as journal articles, software notes, or even the "about" pages of an organization's website, the phylomemetic approach provides an alternative line of evidence to support – or challenge – existing narratives of a community's history. Additionally, understanding how the content or completeness of

https://doi.org/10.23106/dcmi.952137630

metadata records evolve over time can inform the work of metadata creators, and metadata repository managers. For instance, changes in how users create records (e.g. filling in more or less fields, with more or less clarity) can be indicative of larger trends within a community that may need to be addressed by alterations in policy or best practices. A phylomemetic view of metadata collections may help repository managers understand and guide their user communities.

3. Software Package Metadata

Here we demonstrate this approach with an analysis of metadata records describing different packages of the Debian operating system. The Debian operating system is one of the most successful distributions of Linux – a free open source software alternative to commercial operating systems, such as Windows and Mac OS. Each new distribution of Debian contains over four hundred individual different software packages. For instance, just as each distribution of Windows comes with a word processing package (e.g. Microsoft® Word) so too does Debian (e.g. AbiWord). The different package configurations of a distribution represent significant changes in the people and the politics of an open-source project as an institution. While these changes are described in the software documentation, our phylomemetic analysis will provide us with an alternative line of evidence, through which we may better understand the changes of this software, and its development community, over time.

The workflow we have developed is as follows:

- We harvest descriptive metadata records about different software packages found in each Debian distribution (e.g. word processing software packages). Each package's metadata are coded to create a character matrix.
- A package's character matrix represents differences or changes in a package over time - collectively the different package matrixes represent the 'genetic makeup' of a Debian distribution. This is much like a biologist would compare individual characteristics of one specimen to another and code for absence or presence of common features.
- We then load this matrix into phylogenetic software - PAUP (Swofford, D. L., & Begle, 2013) - to produce a visualization of the different Debian distributions.
- We set PAUP to use a maximum likelihood algorithm - which sorts characteristics by their relevant distance (difference) from one another.
- PAUP then produces a tree' that visualizes the relevant divergence of each distribution.
- The tree can then be used to infer differences in how package configurations represent differences in Debian distributions, potentially revealing substantive changes in the institutional features of the broader open-source project.

4. Future Work

Phylomemetic studies of metadata collections are related to a number of previous evolutionary studies in knowledge representation and classification research. For instance, work by Krause et al (2015) and Tennis (2002, 2012) is of particular relevance to the modeling of cultural change using metadata as a primary source. We follow these authors in noting that metadata creation methods evolve just as much as the artifacts they describe; thus, a phylomemetic analysis is potentially a way to not just study the relatedness and evolution of records, but the evolution of different methods of metadata application development and design.

References

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology*, *54*(4), 513-526.

Howe CJ, Windram HF (2011) Phylomemetics—Evolutionary Analysis beyond the Gene. PLoS Biol 9(5): e1001069. doi: 10.1371/journal.pbio.1001069

https://doi.org/10.23106/dcmi.952137630

Krause, E. M., Clary, E., & Greenberg, J. (2015). Evolution of an Application Profile: Advancing Metadata Best Practices through the Dryad Data Repository. In International Conference on Dublin Core and Metadata Applications (pp. 63-75).

Le Bomin, S., Lecointre, G., & Heyer, E. (2016). The Evolution of Musical Diversity: The Key Role of Vertical Transmission. *PloS one*, *11*(3), e0151570.

O'Brien, M. J., Darwent, J., & Lyman, R. L. (2001). Cladistics Is Useful for Reconstructing Archaeological Phylogenies: Palaeoindian Points from the Southeastern United States. *Journal of Archaeological Science*, *28*(10), 1115–1136. doi:10.1006/jasc.2001.0681

Swofford, D. L., & Begle, D. P. (1993). *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1, March 1993*. Center for Biodiversity, Illinois Natural History Survey.

Tëmkin, I., & Eldredge, N. (2007). Phylogenetics and material cultural evolution. Current Anthropology, 48(1), 146-154.

Tennis, J. T. (2002). López-Huertas, M. (ed.) Subject Ontogeny: Subject Access through Time and the Dimensionality of Classification. In Challenges in Knowledge Representation and Organization for the 21st Century: Integration of Knowledge across Boundaries: Proceedings of the Seventh International ISKO Conference. Vol. 8. 54 - 59. Ergon Verlag. Würzburg.

Tennis, J. T. (2012). The strange case of eugenics: A subject's ontogeny in a long-lived classification scheme and the question of collocative integrity. Journal of the American Society for Information Science and Technology, 63(7), 1350-1359.

Tehrani, J. J. (2013). The phylogeny of little red riding hood. *PloS one*, *8*(11), e78871.

Thomer, A. K., & Weber, N. M. (2014). The phylogeny of a dataset. Proceedings of the American Society for Information Science and Technology, 51(1), 1-11.

Underwood, T. (2016) The Lifecycle of Genres. Journal of Cultural Analytics. 1(1). Retrieved from: http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/

https://doi.org/10.23106/dcmi.952137630