The Dryad Repository: Designing a Curation Workflow

Jane Greenberg[1], Hilmar Lapp[2], Ryan Scherle[2], Todd Vision[2], Hollie White[1], Sarah Carrier[1], Peggy Schaeffer[2]

[1] School of Information and Library Science, University of North Carolina, USA

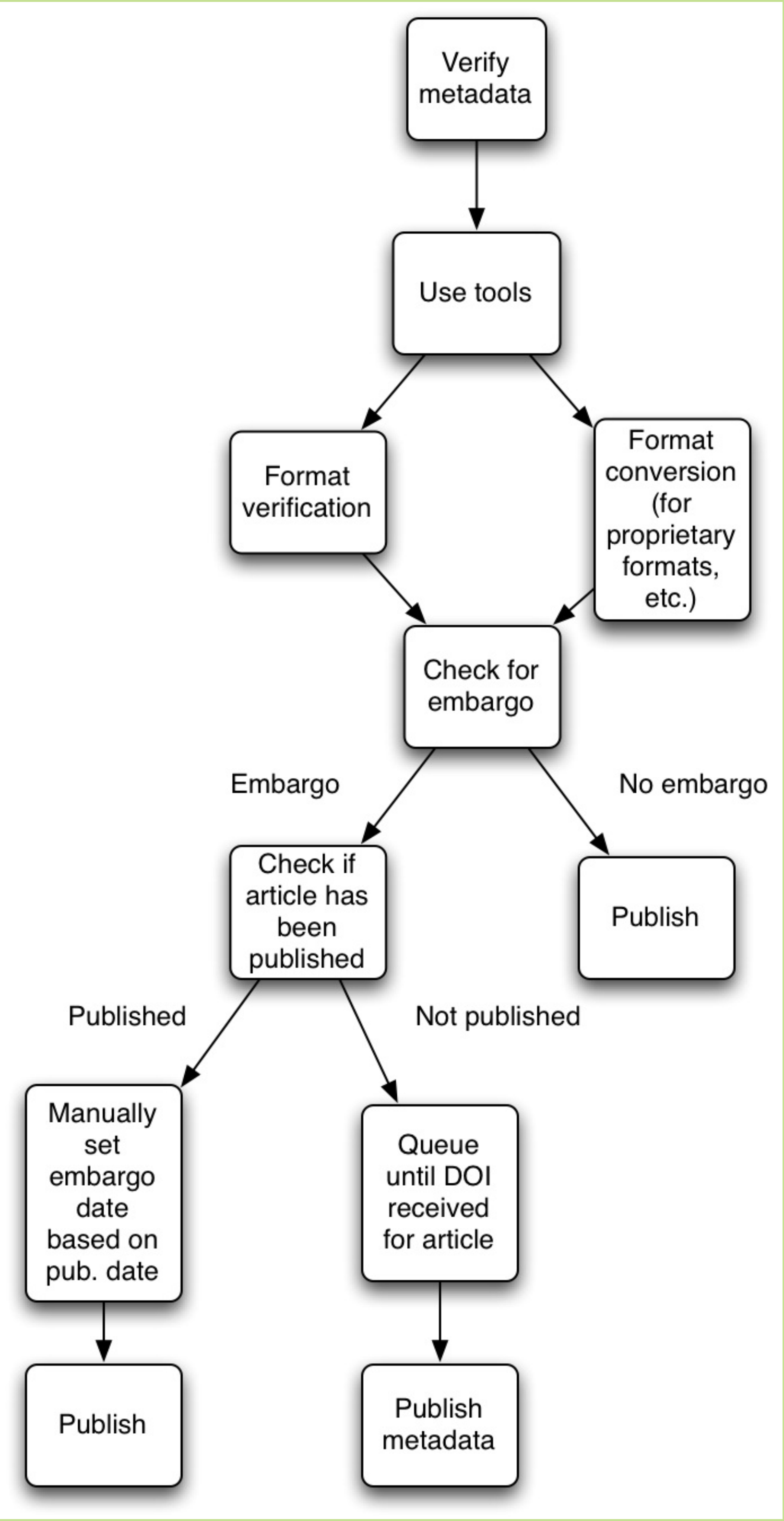
[2] National Evolutionary Synthesis Center (NESCent), USA

Summary

Dryad is a repository of data underlying scientific publications, with an initial focus on evolution, ecology, and related fields. When an author publishes an article, some types of supporting data are deposited in well-known archives like GenBank and TreeBASE, but other types of data have no permanent home. Dryad provides that home.

Dryad's curation workflow integrates automatic and human metadata generation techniques and leverages depositor, scientist and professional curator expertise. The curation workflow has been informed by results from a survey involving 400 prospective Dryad depositors, intensive semi-structured interviews with 17 evolutionary biologists (Carrier, 2008; White, 2008), a metadata content analysis of eight schemes (Greenberg, 2009), a vocabulary mapping study including nearly 600 terms, and stakeholder feedback.

Data Curation Workflow


Verify metadata
Use tools
Format verification
Format conversion (for proprietary formats, etc.)
Check for embargo
Embargo
No embargo
Check if article has been published
Publish
Published
Not published
Manually set embargo date based on pub. date
Queue until DOI received for article
Publish
Publish metadata

Vocabulary Analysis

A vocabulary assessment was conducted to verify identify appropriate vocabularies for representing Dryad data objects.

- A sample of approximately 600 keywords was collected from 104 articles appearing in selected issues of five partner journals (*American Naturalist*, *Molecular Biology and Evolution*, *Systematics Biology*, *Molecular Ecology*, and *Evolution*).
- Terms were categorized into nine facets (topic, research method, geographic location, taxon, personal name, agency name, anatomical aspect, discipline, and habitat), and searched for in appropriate vocabulary sources (e.g., *NBII Thesaurus*, *LCSH*, *Getty Thesaurus of Geographic Names* (TGN), *Gene Ontology* (GO) -- to name a few).
- Search metrics gathered scores for exact match, partial match, and no match.
- No single vocabulary was found sufficient for Dryad, but portions of existing vocabularies were shown to be valuable.
- *Example 1: 431 terms searched in the NBII Thesaurus, 25% of the terms were exact matches, while 75% were partial and non-matches.*
- *Example 2: 531 terms were searched in LCSH, with 22% found to be exact matches and 78% partial and non-matches.*

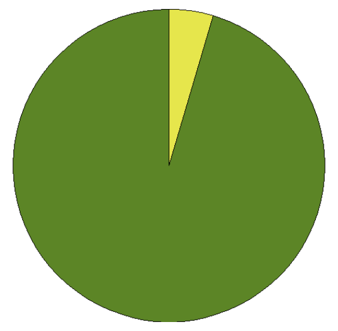
Selected results of this project are reported on in Greenberg (2009), and provided evidence for pursuing the Helping Interdisciplinary Vocabulary Engineers (HIVE) project (https://www.nescent.org/sites/hive/Main_Page), supported by Institute of Museum and Library Services.

Conclusion

The curation process needs access to a number of different vocabularies.

Survey

400 evolutionary biologists were surveyed from EvolDir to understand their interactions with existing data archives, their data sharing practices, and their dependency on digital media for research and reporting.



- 95% of respondents think the data underlying published scientific results should be made publicly accessible.
- 24% have reused previously published data that was at least 10 years old.
- 38% chose "submission to database" as the most preferred method of providing data to others.
- 66% report having received at least one request to share data from published works.

Interviews

Two small-scale intensive interview studies were conducted, by White (2008) with 7 participants, and Carrier (2008), with 10 participants. These studies provide insight into the current data curation and sharing practices of evolutionary biologists.

- Research data is stored in local databases and hard drives using a variety of organization schemes.
- Excel spreadsheets and images comprise much collected data.
- Respondents were generally open to sharing data, though many still want to retain control over how the data will be used.
- Metadata and organization are recognized as being important.

References

- Carrier, S. (2008). Unpublished raw data.
- Greenberg, J. (2009). Theoretical considerations of lifecycle modeling: an analysis of the Dryad repository demonstrating automatic metadata propagation, inheritance, and value system adoption. *Cataloging & Classification Quarterly*, 47, 3: 380-402.
- White, H.C. (2008). Exploring evolutionary biologists' use and perceptions of semantic metadata for data curation. *International Conference on Dublin Core and Metadata Applications. Berlin, Germany.*

Poster designed by Mike Graves.