

# **SUBJECT TAGGING: RECOMMENDATIONS FOR DRYAD CURATORS AND SCIENTISTS**

**Priscilla Jane Frazier**

14 April, 2013

SUBJECT TAGGING: RECOMMENDATIONS FOR DRYAD CURATORS AND SCIENTISTS	1
PROBLEM SPECIFICATION	3
THE VALUE OF QUALITY DESCRIPTIVE METADATA	4
METRICS FOR METADATA QUALITY	4
HUMAN METADATA GENERATION	5
MODELS FOR CREATING METADATA	6
CONTROLLED TERMS	7
WHAT IS A CONTROLLED VOCABULARY?	7
THE VALUE OF CONTROLLED TERMS	7
CONTROLLED VOCABULARIES USEFUL IN DESCRIBING DRYAD DATA	7
UNCONTROLLED TERMS	10
THE VALUE OF UNCONTROLLED TERMS	11
FOLKSONOMIES	11
RECOMMENDATIONS	12
HOW TO USE THE DESCRIPTIVE METADATA FIELDS	12
SUBJECT KEYWORDS	13
TEMPORAL KEYWORDS	14
SPATIAL KEYWORDS	15
TAXONOMIC KEYWORDS	16
BIBLIOGRAPHY	19
APPENDIX 1	22
FURTHER RESOURCES FOR AUTHORS	22

## PROBLEM SPECIFICATION

The Dryad project is an online repository for the data which underlie publications in the biosciences. Submissions to Dryad consist of the data that is used to create scientific publications, for example phylogenetic trees, tables, spreadsheets, images, maps, gene alignments, matrices, and the like. When authors submit their data to Dryad, they have the opportunity to enter topical headings/subject headings that will be used to categorise and retrieve their data in the future.

There are currently four types of topical metadata that Dryad accepts: subject, temporal, spatial, and taxonomic. Because these headings are not controlled in any way, scientists often do not submit any headings (leave the fields blank) or submit them in ways that might not benefit future users of Dryad (poor formatting, abbreviations, etc.). The data is much more accessible to users of Dryad - other scientists, the public, Dryad curators - if the data is described in a meaningful and thorough way.

Studying this topic has allowed me to gain a great understanding of how scientists tag their data. Do they come up with their own topical terms, do they consult controlled vocabularies, or do they use some other source for these terms? The goal of this project will be to write a memo or guide which will help Dryad librarians guide scientists to describe their data in the best way possible by submitting meaningful subject headings/topical headings with their data. A summary of the benefits of scientific data archival is best summed up in Whitlock's 2009 article *Data archiving in ecology and evolution: best practices*:

“Data archives serve science in a variety of ways. Publicly archived data enable more transparent science, with better error checking and verification of results. Archiving also enables data to be re-used for broader meta-analyses and to address new questions. Available data can serve a powerful educational role, both in teaching the statistical and technical aspects of research and to engage students in the process of science. Public data archiving is also a powerful mechanism for data security, providing a mechanism by which data can be saved and re-accessed by the original authors and others even after hard disk failure or other catastrophes.”

# THE VALUE OF QUALITY DESCRIPTIVE METADATA

---

## METRICS FOR METADATA QUALITY

Metadata are defined as the data providing information about aspects of the data. This may include the means of the creation of data, the purpose of the data, the time and date of the creation, the name of the creator or author of the data, the location where the data was created, or descriptions of the data's *about-ness*.

There has been much research on criteria which can and should be used to measure the quality of this metadata. Rotherberg (1996) identifies correctness and appropriateness as two main criteria for data evaluation. In their 1997 article *The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): Results from an exploratory study*, Moen et al. describe 23 evaluation criteria with which to analyse metadata quality. From these 23 criteria, the researchers distilled four main criteria: accuracy, consistency, completeness, and currency. These four criteria partially overlap with Tozer's (1999) data quality measures of accuracy, consistency, completeness, timeliness, and intelligibility. Bruce and Hillman (2004) refine the previously mentioned criteria and modify them for the library community, suggesting completeness, accuracy, provenance, conformance to expectation, logical consistency, coherence, timeliness, and accessibility. In her 2009 article, Park evaluates Moen et al. (1997) and Bruce and Hillman's (2004) criteria in addition to the criteria determined by seven other research teams, and determines that accuracy, completeness, and consistency are the most commonly used criteria in measuring metadata quality.

Accuracy (or correctness) of metadata indicates the accurate description and input of data. According to Park (2009), this is made up of three elements: the accurateness of the content of the data element, the correctness of the intellectual property, and the correctness of the instantiation (or particular instance). Park includes errors in spelling, date format, capitalisation, punctuation as well as non-authoritative forms of terms, typographical errors and incorrect data values as potential problems with the accuracy of metadata.

According to Park, completeness can be measured by full access capacity to individual resources and connection to the collections in which they are housed. Completeness points to resource discovery as the functional purpose of metadata, but

does not necessarily indicate that all elements in a particular metadata scheme must be used. Because completeness is directly affected by policies, best practices, and application profiles for specific domains, the completeness of a metadata records may vary depending on the environment in which they are housed. Completeness can be achieved in a metadata record if the given resource type, its relation to the local collection and the local metadata guidelines are met satisfactorily.

Consistency (or comparability) can be measured by examining the values of the metadata and the format of the metadata. Park states that metadata values must be examined on the conceptual level by measuring the degree to which the same data values are used for delivering similar concepts in the description of a resource, and the data format must be examined on a structural level by measuring the extent to which the same structure or format is used for presenting similar attributes of a resource. For example, differences between the encoding of a date element (e.g., MM-DD-YYYY versus DD-MM-YY) are structurally inconsistent, and may cause problems for future users of the data.

---

## HUMAN METADATA GENERATION

According to Greenberg and Robertson (2002), human metadata generation takes place when an individual is responsible for the identification and assignment or recording of resource materials. This type of metadata generation may take place in a number of ways by different types of individuals. The three types of individuals that have been identified in the literature are professional metadata creators, resource authors, and social taggers. Professional metadata creators may be catalogers, indexers or curators who have formal training and are proficient in the use of descriptive standards (Greenberg et al. 2002). According to Lu et al. (2002), social taggers apply their own descriptors to sources that interest them. Resource authors, like the submitters of data to Dryad, are the individuals responsible for the creation of the intellectual content of a work. Most importantly, they are “intimate with their creations and have knowledge of unrecorded information for producing descriptive metadata,” allowing them to have a unique ability to describe their data with the highest accuracy (Greenberg and Robertson 2002).

The fact that resource authors may lack the knowledge of indexing and cataloging principles that professional metadata creators possess is well documented (Greenberg et al. 2003). However, attitudes toward resource authors as metadata creators are somewhat split in the literature. Wilson, in her 2007 article *Toward Releasing the Metadata Bottleneck*

states that resource authors “seldom provide sufficient metadata for their digital resources” and Greenberg et al. (2003) state that authors have “reported confusion or uncertainty regarding specific fields and have requested greater assistance in determining appropriate inputs, especially for subject fields.” However, Greenberg et al. also report that resource authors state a desire for better understanding of the metadata record and its purpose. Currier et al. (2003) discuss the debate between allowing metadata professionals or resource authors to create metadata for resources:

“How may this difficult and complex task best be carried out for maximum resource discoverability by a heterogeneous population of searchers? Should the resource author, who may know their subject area and its terminology well, create the subject metadata? Or should it be a metadata specialist, who may know the specific area less well, but may be better placed to step back and think about all the potential users of a resource, and about consistency of key words and classifications across a repository or network?”

---

## MODELS FOR CREATING METADATA

Currier et al. (2003) recommend three models for metadata creation: creation by a resource author only, creation by a metadata specialist only, or creation by collaboration between a resource author and a metadata specialist. A data collection centre (like Dryad) would be well advised to ensure that their system supports user support and training if they are to rely on metadata creation by resource authors alone. Conversely, metadata specialists, who already possess the skills needed to create a quality metadata record may lack the knowledge about the context, history or subject area of the resource in order to best record its metadata. Currently, Dryad uses a semi-collaborative approach to metadata creation. Although resource authors do not consult with the curators, or metadata specialists while they are submitting data, the curators spend time checking the author-created metadata for quality issues. Greenberg and Robertson (2002) recommend this model, stating that “...the integration of expert and author generated descriptive metadata can advance and improve the quality of metadata for web content, which in turn could provide useful data for intelligent web agents, ultimately supporting the development of the Semantic Web. [...] If such partnerships are well planned and evaluated, they could make a significant contribution to achieving the Semantic Web.” Along with the three models for metadata creation that Currier et al. (2003) present, social tagging and the rise of

folksonomies should be mentioned as a fourth model. Folksonomies will be discussed in a later section of this paper.

## CONTROLLED TERMS

---

### WHAT IS A CONTROLLED VOCABULARY?

A controlled vocabulary allows for organisation of some content, or knowledge, in a way in which it can be easily retrieved at a later time. Vocabularies are 'controlled' in that they make use of authorised descriptions of the content they contain. These groupings of concepts are carefully selected and described so that the information they contain can be retrieved in the most efficient ways possible.

---

### THE VALUE OF CONTROLLED TERMS

Natural language, or the way that humans speak in everyday life, is messy. We use multiple terms and phrases to describe the same things, and there are fine (or grey) lines between one meaning and another. The organisation, categorisation, and labelling of our knowledge can be achieved by way of controlled vocabularies. A controlled vocabulary allows for all concepts to be consistently labeled using language that is unambiguous and is familiar to its users. More importantly, controlled vocabularies allow us to search for concepts and achieve successful, quality results.

---

### CONTROLLED VOCABULARIES USEFUL IN DESCRIBING DRYAD DATA

**Medical Subject Headings<sup>1</sup>** (MeSH) is the controlled vocabulary of the United States National Library of Medicine. Currently consisting of more than 177,000 terms situated in a twelve-level hierarchy, MeSH allows for the indexing of articles from biomedical journals for the MEDLINE/PubMED database. MeSH is comprised of three main types of terms: descriptors (main headings), qualifiers (subheadings), and supplementary concept records (SCRs). Descriptors indicate the subject of citations indexed in MEDLINE/PubMED, and the 83 existing topical qualifiers allow for the grouping

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/mesh>

together of citations concerned with a particular aspect of a subject. SCRs index chemicals and drugs and are searchable by substance name in PubMed.

The Getty **Thesaurus of Geographic Names**<sup>2</sup> (TGN) is a controlled vocabulary provided by the Getty Vocabulary Program of the J. Paul Getty Trust. TGN currently includes approximately 1,106,000 hierarchically arranged terms that describe names and associated information about places, including current and historical physical features and political entities. Each term entry includes a unique identification number, known as a subject ID, text description about the place, geographical coordinates, associated place-names, dates referring to the usage of those names, position of the entry in the TGN hierarchy, information about related places, information about the type of place described in the entry, and information about the data source.

Developed by the White House Subcommittee on Biodiversity and Ecosystem Dynamics, the **Integrated Taxonomic Information System**<sup>3</sup> (ITIS) is a controlled vocabulary for describing ecosystem management and biodiversity conservation. Information about each species includes an authoritative scientific title, a taxonomic rank and serial number, associated synonyms and vernacular names, information about the data source, and data quality indicators.

The **BIOSIS Controlled Vocabulary** contains multiple lists of terms used in the BIOSIS Previews and Biological Abstracts databases. The vocabulary is organised into several categories including concepts, organism classifiers, and geopolitical locations, among others. The vocabulary includes 168 “major concepts” and 562 “concept codes” used for subject or topical indexing; 77 “organism classifiers” and 957 “super taxa” used for taxonomic data; and 316 geopolitical locations.

The National Biological Information Infrastructure (NBII) was a program coordinated by the United States Geological Survey's Biological Informatics Program Office<sup>4</sup>. Its purpose was to facilitate access to data and information on the biological resources of the United States, utilising government agencies, academic institutions, non-government organisations, and private industry. The **NBII Biocomplexity Thesaurus**, and online thesaurus of scientifically reviewed biological terms, was initially created through a merger

---

<sup>2</sup> <http://www.getty.edu/vow/TGNSearchPage.jsp>

<sup>3</sup> <http://www.itis.gov/>

<sup>4</sup> Development and web hosting of the NBII was terminated 15 January 2012.



of several individual thesauri, including the CSA Aquatic Sciences and Fisheries Thesaurus, the Cambridge Scientific Abstracts (CSA) Life Sciences Thesaurus, the CSA Pollution Thesaurus, the CSA Sociological Thesaurus, the CERES/NBII Thesaurus, and the CSA Ecotourism Thesaurus. The thesaurus includes over 15,000 terms on subjects such as aquatic sciences, life sciences, social sciences, ecotourism, and pollution.

The **Library of Congress Subject Headings**<sup>5</sup> (LCSH) is a controlled vocabulary for use in subject cataloging and indexing. First published in 1898, LCSH was designed for and is maintained by the Library of Congress, but the system has been adopted by many other libraries. LCSH covers all subjects generally. Subject headings can consist of single words or phrases and are divided into two types: main headings and subheadings. LCSH uses four categories of subdivisions to further distinguish main heading topics: form subdivisions, geographical subdivisions, chronological subdivisions, and topical subdivisions.

**AGROVOC**<sup>6</sup> is a multilingual controlled vocabulary covering all areas of interest to the Food and Agricultural Organisation of the United Nations (FAO), including food, nutrition, agriculture, fisheries, forestry, and the environment. AGROVOC contains over 30,000 concepts organised in a hierarchy, and concepts may have labels in up to 22 languages.

Wilson and Reeder's **Mammal Species of the World**<sup>7</sup> is an online database of mammalian taxonomy. Use of the Mammal Species of the World, through search or taxonomic browsing, allows users to verify recognised scientific names and conduct taxonomic research.

**CAB Thesaurus**<sup>8</sup> is the research tool for users of the CAB ABSTRACTS™ and Global Health databases. The thesaurus includes over 200,000 terms broad and covers topics in the applied life sciences, technology and social sciences.

---

<sup>5</sup> <http://id.loc.gov/authorities/subjects.html>

<sup>6</sup> <http://aims.fao.org/standards/agrovoc/functionalities/search>

<sup>7</sup> <http://www.vertebrates.si.edu/msw/mswcfapp/msw/index.cfm>

<sup>8</sup> <http://www.cabi.org/cabthesaurus/>

The **GeoRef Thesaurus**<sup>9</sup> contains 23,065 valid and 7,740 invalid terms, of which about 1780 are newly added. The Thesaurus is a guide to the index terms used in GeoRef, a database consisting of bibliographic citations and abstracts covering the field of geology and its allied environmental sciences. For each term, the Thesaurus includes hierarchical and other relationships, usage notes, dates of addition, indexing rules, geographic coordinates, and guidelines for searching. Cross-references from invalid to valid terms are included.

The National Agricultural Library's **NAL Agricultural Thesaurus**<sup>10</sup> includes terminology which supports biological, physical and social sciences. Biological nomenclature comprises a majority of the terms in the thesaurus and is located in the "Taxonomic Classification of Organisms" Subject Category. Political geography is also included, and is mainly described at the country level.

**uBIO**<sup>11</sup> is an initiative within the science library community to join international efforts to create and utilise a comprehensive and collaborative catalog of known names of all living (and once-living) organisms. uBio's Taxonomic Name Server (TNS) catalogs names and classifications to enable tools that can help users find information on living things using any of the names that may be related to an organism.

## UNCONTROLLED TERMS

Uncontrolled terms, or tags, are taken directly from natural language. Because they do not possess the same characteristics of controlled vocabulary terms, they pose many advantages – and disadvantages – over the use of controlled vocabulary terms in the submission of data to Dryad.

---

<sup>9</sup> <http://www.agiweb.org/georef/lists.html>

The entire GeoRef Thesaurus may also be found in this PDF document:  
<http://www.agiweb.org/georef/PDF/Introduction.pdf>

<sup>10</sup> [http://agclass.nal.usda.gov/dne/search\\_sc.shtml](http://agclass.nal.usda.gov/dne/search_sc.shtml)

<sup>11</sup> <http://www.ubio.org/>

---

## THE VALUE OF UNCONTROLLED TERMS

According to Noruzi (2006), uncontrolled terms, or tags, are words or phrases users attach to resources that may help in later retrieval of that resource. These tags have no fixed categories, syntaxes, or standards. However, the fact that no time was taken to develop standards or categorisations for these tags means that there is little overhead in the efficiency of their creation. They are created precisely at the point of submission, and potentially cost little time and effort on the part of the user, be it a resource author or a social tagger. Lu et al. (2010) present several compelling advantages to the use of tags. First, they state that tags may help to bridge the gap between professional and public discourse by providing a source of terms not included in controlled vocabularies. Second, they mention that tags not only allow users to search resources in their own language, but also provide a window for the libraries to understand and learn more about user information needs and interests. Third, their findings show that social taggers may help enhance subject access to collections by describing resources with terms different from those used by experts (this final sentiment is echoed in Rolla 2009).

Along with these many advantages, uncontrolled terms also pose several important and challenging disadvantages. Because tags are not controlled in any way, a certain individuals' tags may conflict with another individuals' tags. These conflicts may manifest themselves as polysemy (words that have several meanings), synonymy (different words with similar or identical meanings), plurality (inconsistencies in the use of plurals), or granularity (inconsistencies in the depth or specificity of tags). Any of these problems may lead to low precision in searching.

---

## FOLKSONOMIES

A portmanteau of the words *folks* and *taxonomy*, folksonomy is an internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels that categorise content such as web resource, online photographs, and web links (Noruzi 2009). Folksonomies are created by social taggers, not information professionals, and these taggers assign one or more tags to each resource for their own individual use which is then shared through a community.

Much research has been done to study the use of user tags versus the use of controlled vocabulary terms. Lu et al. (2010) found that only a fraction of tag vocabulary terms overlap with LCSH terms, and even those overlapping terms might be used by social taggers and information professionals in different ways. Rolla (2009) reports that users of the LibraryThing<sup>12</sup> social cataloging web application assign tags that range in depth from general to specific, whereas LCSH terms assigned to corresponding bibliographic records are more general in nature. In addition, cataloger-assigned LCSH terms in approximately 55% of bibliographic records brought out topics or concepts that LibraryThing tags did not, and approximately 75% of the time catalogers and taggers agreed on at least a portion of what a book is 'about'. In conclusion, the Library of Congress Working Group on the Future of Bibliographic Control reports in 2008 that "allowing user-supplied data in online catalogs will make the catalogs more relevant to users accustomed to the internet and also will improve access to the materials in the library collection."<sup>13</sup>

## RECOMMENDATIONS

---

### HOW TO USE THE DESCRIPTIVE METADATA FIELDS

In recommending best practices for providing terms for the descriptive metadata fields in Dryad, the author would urge authors to consider the accuracy, consistency, and completeness of the chosen terms first before submission. In addition, the use controlled vocabulary terms are suggested, but not required. The author is recommended to weigh the benefits and disadvantages of submitting their own tags versus authorised controlled vocabulary terms. The following sections highlight the four individual keyword fields, and give specific instructions on how to best provide terms.

---

<sup>12</sup> <http://www.librarything.com/>

<sup>13</sup> Library of Congress Working Group on the Future of Bibliographic Control (2008) On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control.

---

## SUBJECT KEYWORDS

The image shows a web form for entering subject keywords. A yellow callout box contains the following text: "Please enter general keywords associated with the data file. Keywords may be separated by commas, or added individually. For example: adaptation evolutionary contingency founder effects". The form includes a "Subject keywords:" field with an "Add" button, a "Taxonomic names:" field, a "Geographic areas covered:" field, and a "Geologic timespans covered by this publication:" field with an "Add" button. At the bottom, there are "Save & Exit" and "Continue to describe data file" buttons.

Submitters of data to Dryad are required to include at least one subject keyword with their data submission. Within the Dryad submission system this field is repeatable, meaning that a submitter may include as many subject keywords as they choose. The submission of multiple subject keywords may be achieved by separating individual keywords by a comma. Using semicolons, dashes, periods, or any other types of punctuation will result in a list of keywords concatenated into single keyword. For example, if the two subject keywords *Facial structure* and *Testosterone* are entered into the subject field as [Facial Structure; Testosterone] the two will be recognised in Dryad’s system as a single entity and will be shown as one subject keyword “Facial structure; Testosterone”, not as two separate subject keywords “Facial structure” and “Testosterone.”

According to Dryad’s metadata schema, the subject keyword field is associated with the Dublin Core metadata term *Subject*.<sup>14</sup> According to the Dublin Core Metadata Initiative (DCMI), a subject “will be represented using keywords, key phrases, or classification codes.” DCMI also recommends the use of controlled vocabulary terms for use in the *Subject* field. Because Dryad does not support the use of integrated controlled vocabularies, data submitters may either create their own subject keywords or they may draw from any controlled vocabulary they choose.

## RECOMMENDED CONTROLLED VOCABULARIES

The following controlled vocabularies are recommended for reference in adding subject keywords to Dryad data submissions:

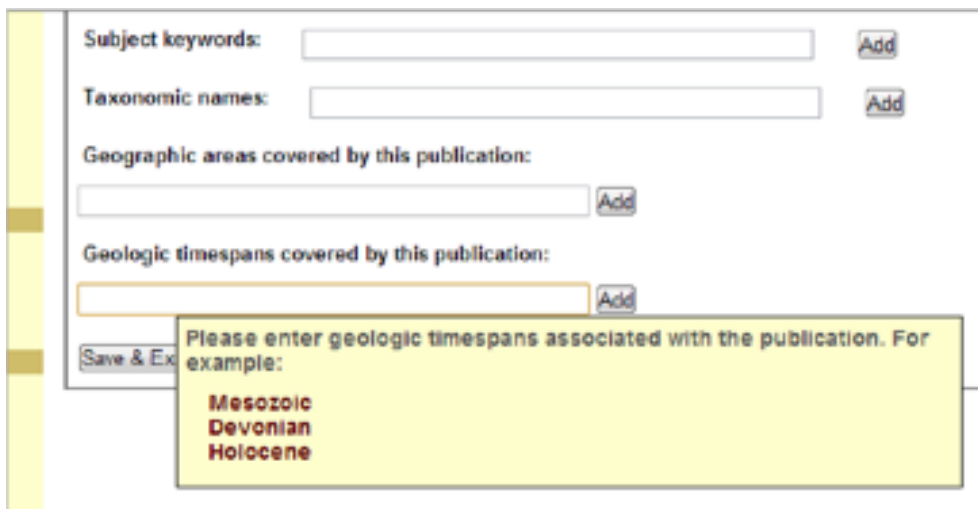
---

<sup>14</sup> DC term *Subject* is located at the namespace <http://purl.org/dc/elements/1.1/subject>

- AGROVOC
- BIOSIS
- CAB Thesaurus
- GeoRef
- LCSH
- Mammal Species of the World
- MeSH
- NAL Agricultural Thesaurus
- uBio

---

## TEMPORAL KEYWORDS



The screenshot shows a web form with several input fields and buttons. The fields are labeled: 'Subject keywords:', 'Taxonomic names:', 'Geographic areas covered by this publication:', and 'Geologic timespans covered by this publication:'. Each field has an 'Add' button to its right. A 'Save & Exit' button is located at the bottom left. A yellow callout box is overlaid on the 'Geologic timespans' field, containing the text: 'Please enter geologic timespans associated with the publication. For example: Mesozoic, Devonian, Holocene'.

Submitters of data to Dryad are not required to include temporal keywords with their data submission, but are urged to do so if the field is applicable to the nature of the data. This field in the submission system is also repeatable, meaning that a submitter may include as many temporal keywords as they choose. Again, the submission of multiple temporal keywords may be achieved by separating individual keywords by a comma, and using semicolons, dashes, periods, or any other types of punctuation will result in a list of keywords concatenated into single keyword.

According to Dryad's metadata schema, the temporal keyword field is associated with the Dublin Core metadata term *Temporal*.<sup>15</sup> According to the DCMI, a temporal keyword should be used to describe "temporal characteristics of the resource."

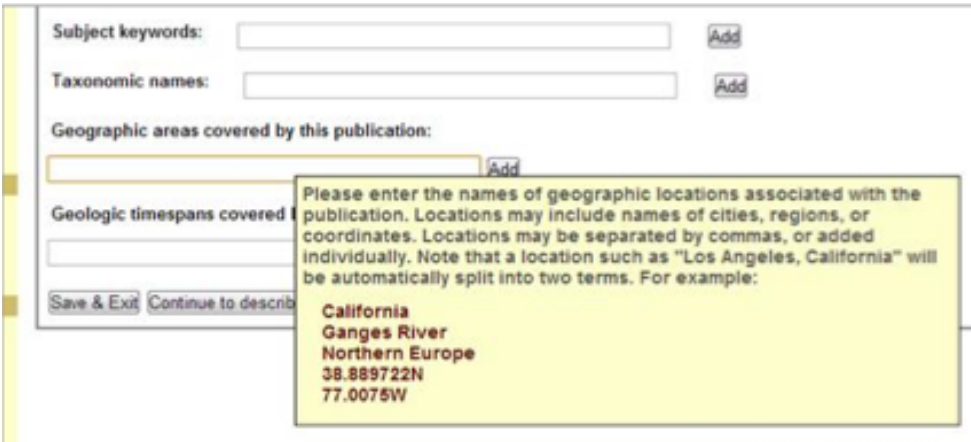
## RECOMMENDED CONTROLLED VOCABULARIES

The following controlled vocabularies are recommended for reference in adding temporal keywords to Dryad data submissions:

- LCSH
- MeSH
- TGN

---

## SPATIAL KEYWORDS



The screenshot shows a web form with several input fields. The 'Geographic areas covered by this publication' field is highlighted with a yellow tooltip. The tooltip text reads: 'Please enter the names of geographic locations associated with the publication. Locations may include names of cities, regions, or coordinates. Locations may be separated by commas, or added individually. Note that a location such as "Los Angeles, California" will be automatically split into two terms. For example: California, Ganges River, Northern Europe, 38.889722N 77.0075W'. Other fields include 'Subject keywords', 'Taxonomic names', and 'Geologic timespans covered'. Buttons for 'Add', 'Save & Exit', and 'Continue to describe' are also visible.

Submitters of data to Dryad are not required to include spatial keywords with their data submission, but are urged to do so if the field is applicable to the nature of the data. This field in the submission system is also repeatable, meaning that a submitter may include as many spatial keywords as they choose. Again, the submission of multiple spatial keywords may be achieved by separating individual keywords by a comma, and using semicolons, dashes, periods, or any other types of punctuation will result in a list of keywords concatenated into single keyword. Submitters of data to Dryad should note that

---

<sup>15</sup> DC term Temporal is located at the namespace <http://purl.org/dc/elements/1.1/temporal>

locations with multi-part names, such as *Los Angeles, California*, will be automatically split into two terms.

According to Dryad's metadata schema, the spatial keyword field is associated with the Dublin Core metadata term *Spatial*.<sup>16</sup> According to the DCMI, a spatial keyword should be used to describe "spatial description of the dataset specified by a geographic description and geographic coordinates." The instructions given for entering data into this field in the Dryad submission system indicate that "locations may include names of cities, regions, or coordinates." Like the DCMI, the Dryad curation team recommends using terms from standard taxonomies of controlled vocabularies for use in the spatial keyword field.

## RECOMMENDED CONTROLLED VOCABULARIES

The following controlled vocabularies are recommended for reference in adding spatial keywords to Dryad data submissions:

- BIOSIS
- LCSH
- TGN

---

## TAXONOMIC KEYWORDS

Submitters of data to Dryad are not required to include taxonomic keywords with their data submission, but are urged to do so if the field is applicable to the nature of the data. This field in the submission system is also repeatable, meaning that a submitter may include as many taxonomic keywords as they choose. Again, the submission of multiple

---

<sup>16</sup> DC term Spatial is located at the namespace <http://purl.org/dc/elements/1.1/spatial>



taxonomic keywords may be achieved by separating individual keywords by a comma, and using semicolons, dashes, periods, or any other types of punctuation will result in a list of keywords concatenated into single keyword.

According to Dryad's metadata schema, the taxonomic keyword field is associated with the Darwin Core metadata term *Specific Epithet*.<sup>17</sup> According to the Biodiversity Information Standards (TDWG), a taxonomic keyword should be used to describe "The specific epithet of the scientific name applied to the organism." The instructions given for entering data into this field in the Dryad submission system indicate that taxonomic keywords should be used to describe "the full name of the lowest level taxon to which the organism has been identified in the most recent accepted determination, specified as precisely as possible." Like the DCMI, the Dryad curation team recommends using terms from standard taxonomies of controlled vocabularies for use in the spatial keyword field. The following quotation, taken from the Borer et al. (2009) article *Some Simple Guidelines for Effective Data Management*, discusses the problems associated with taxonomic keywords and suggested solutions.

"Over time, the names of taxa often are changed as their evolutionary relationships are clarified. The same taxonomic name can actually refer to two or more different concepts of a species. However, scientific names in ecological data are fixed as originally recorded, and so it is critical for long-term preservation to document which taxonomic descriptions were intended by each taxon name used in a data set. This becomes particularly important when comparing species information from data collected at different times, as the names used in the data sets can be ambiguous, which affects calculations of diversity and richness, among other issues. The best way to clarify a taxonomic name is to document the taxonomic authority you are using for the name. For example, *Homo sapiens* Linn. clarifies that the authority for this binomial is Linnaeus. Unfortunately, there are several formats to choose from for specifying taxonomic authority information, but any reference information is better than none."<sup>18</sup>

## RECOMMENDED CONTROLLED VOCABULARIES

The following controlled vocabularies are recommended for reference in adding taxonomic keywords to Dryad data submissions:

---

<sup>17</sup> Darwin Core term Specific Epithet is located at the namespace <http://wiki.tdwg.org/twiki/bin/view/DarwinCore/SpecificEpithet>

<sup>18</sup> Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer M. (2009) Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America* 90(2): 205-214.

- BIOSIS
- ITIS
- LCSH
- Mammal Species of the World
- MeSH
- NAL Agricultural Thesaurus
- uBio

## BIBLIOGRAPHY

- Babinec, M. and Mercer, H. (2009) Introduction: Metadata and digital repositories. *Cataloging & Classification Quarterly*. 47: 209-212.
- Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer M. (2009) Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America* 90(2): 205- 214. Retrieved from <http://www.esajournals.org/doi/pdf/10.1890/0012-9623-90.2.205>
- CAB International. (2013). CAB Thesaurus. Retrieved from <http://www.cabi.org/cabthesaurus/>
- Carrier, Sarah W. (2008) The Dryad Repository Application Profile: Process, Development, and Refinement. A Master's paper for the M.S. in I.S. degree.
- Currier, Sarah and Barton, Jane. (2003) Quality Assurance for Digital Learning Object Repositories: How Should Metadata Be Created? *Communities of Practice. ALT-C 2003 Research Proceedings*.
- Greenberg, Jane and Robertson, W. (2002) Davenport. Semantic Web construction: An inquiry of author's views on collaborative metadata generation. *Proc. Int. Conf. on Dublin Core and Metadata for e-Communities*. 45-52.
- Greenberg, Jane et al. (2002) Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization. *Journal of Digital Information* 2(2).
- Greenberg, Jane et al. (2003) Iterative Design of Metadata Creation Tools for Resource Authors. *DC-2003--Seattle Proceedings*.
- ITIS Integrated Taxonomic Information System. (2013). Retrieved from <http://www.itis.gov/>
- American Geosciences Institute. (2013). GeoRef Thesaurus Lists. Retrieved from <http://www.agiweb.org/georef/lists.html>
- Bates, Marcia J. *Encyclopedia of Library and Information Sciences*. 3rd ed, eds M.J. Bates and M.N. Maack. Boca Raton, FL: CRC Press, 2010. Print.
- Bruce, T.R. and Hillman, D. (2004). *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*. In *Metadata in Practice*, eds. D. Hillman and E.L. Westbrook (Chicago: American Library Association).
- Food and Agriculture Organization of the United Nations. (2013). AGROVOC. Retrieved from <http://aims.fao.org/standards/agrovoc/about>
- J. Paul Getty Trust . (2013). *Getty Thesaurus of Geographic Names® Online*. Retrieved from <http://www.getty.edu/vow/TGNSearchPage.jsp>
- Library of Congress. (2013). *Library of Congress Subject Headings*. Retrieved from <http://id.loc.gov/authorities/subjects.html>
- Library of Congress Working Group on the Future of Bibliographic Control (2008) *On the Record: Report of the Library of Congress Working Group on the Future of*

- Bibliographic Control. Retrieved from [www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf](http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf)
- Lu, C., Park, J. and Hu, X. (2010) User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science* 36(6): 763-779.
- Moen, W.E., Stewart, E.L., and McClure, C.R. (1997) The Role of Content Analysis in Evaluating Metadata for the US Government Information Locator Service (GILS): results from an exploratory study.
- Noruzi, Alireza. (2006) Folksonomies: (Un)Controlled Vocabulary? *Knowledge Organization* 33(4): 199-203.
- Page, Roderic. (2006) Taxonomic names, metadata, and the Semantic Web. *Biodiversity Informatics* 3.
- Park, Jung-Ran. (2009) Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly* 47: 213-228.
- Rolla, Peter J.(2008) User tags versus subject headings: Can user-supplied data improve subject access to library collections? *Library Resources & Technical Services* 53(3): 174-184.
- Rothenberg, J. (1996) Metadata to Support Data Quality and Longevity. 1st IEEE Metadata Conference, Silver Spring, Maryland.
- Strader, C. Rockelle. (2009) Author-Assigned Keywords versus Library of Congress Subject Headings. *Library Resources & Technical Services* 53(4): 243-250.
- Tozer, G. (1999) *Metadata Management for Information Control and Business Success*. Boston: Artech House.
- Trant, Jennifer. (2009) Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information* 10(1).
- Marine Biological Laboratory. (2013). uBio Universal Biological Indexer and Organizer. Retrieved from <http://www.ubio.org/>
- National Center for Biotechnology Information, U.S. National Library of Medicine. (2013). MeSH. Retrieved from <http://www.ncbi.nlm.nih.gov/mesh>
- Smithsonian Institution. (2013). Wilson & Reeder's Mammal Species of the World. Retrieved from <http://www.vertebrates.si.edu/msw/mswcfapp/msw/index.cfm>
- Thompson Reuters. (2013). BIOSIS. Retrieved from [http://thomsonreuters.com/products\\_services/science/science\\_products/a-z/biosis/](http://thomsonreuters.com/products_services/science/science_products/a-z/biosis/)
- U.S. Department of Agriculture National Agricultural Library. (2013) Thesaurus: Browse by Subject Category. Retrieved from [http://agclass.nal.usda.gov/dne/search\\_sc.shtml](http://agclass.nal.usda.gov/dne/search_sc.shtml)
- Whitlock, M.C. (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution* 26(2): 61–65. Retrieved from <http://dx.doi.org/10.1016/j.tree.2010.11.006>

Wilson, A.J. (2007) Toward Releasing the Metadata Bottleneck: A Baseline Evaluation of Contributor-supplied Metadata. *Library Resources & Technical Services* 51(1): 16-27.

## APPENDIX 1

---

### FURTHER RESOURCES FOR AUTHORS

The following resources are intended to assist authors in further assistance with the use of controlled and uncontrolled terms in Dryad.

#### **HIVE Browser**

<http://hive.nescent.org/ConceptBrowser.html>

#### **LibraryThing Concepts**

<http://www.librarything.com/concepts>

#### **The Encyclopaedia of Life**

<http://eol.org/>

#### **Depositing data in Dryad frequently asked questions**

<http://datadryad.org/pages/depositing>

#### **How to submit data in Dryad (Youtube video)**

[http://www.youtube.com/watch?feature=player\\_embedded&v=RP33cl8tL28](http://www.youtube.com/watch?feature=player_embedded&v=RP33cl8tL28)