Chelcie Juliet Rowell. Controlled Vocabulary Use by Data Repositories: Determining Status and Potential for Promoting Interoperability. A Master's paper for the M.S. in Information Science degree. July, 2013. 65 pages. Advisor: Jane Greenberg

Controlled vocabularies facilitate interoperability but present challenges related to cost, usability, and interdisciplinarity. The Helping Interdisciplinary Vocabulary Engineering (HIVE) project aims to meet some of these challenges by providing an approach for integrating multiple controlled vocabularies. A companion effort to the ongoing development of HIVE, this research study developed and implemented a Web survey targeting many roles associated with data repositories – data contributors, data curators, DataNet administrators, and repository developers – regarding their uses of controlled vocabularies. Results indicate that a long tail of controlled vocabularies is currently in use by data repositories. Although the convenience sample of this study cannot be generalized to the broader population of data repository stakeholders, the results of this study indicate that a future study could reasonably hypothesize that demand for HIVE-like services exists among data contributors, data curators, and repository developers.

Headings:

Digital libraries

Metadata

Surveys

Thesauri

CONTROLLED VOCABULARY USE BY DATA REPOSITORIES:
DETERMINING STATUS AND POTENTIAL FOR PROMOTING INTEROPERABILTY

by
Chelcie Juliet Rowell

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

July, 2013

Approved by:

Jane Greenberg

Acknowledgements

I would like to express my gratitude to the people who supported the design and implementation of this research study. Robert Losee as well as numerous pilot testers provided valuable feedback on early drafts of the survey instrument. Rebecca Koskela, Laura Moyers, and Amber Budden of DataONE and Mary Whitton of RENCI were instrumental in helping to disseminate the survey. Jane Greenberg served as my faculty advisor. Jane introduced me to the world of metadata research, and working with her on this study has been the deepest learning experience of my time at the School of Information and Library Science of the University of North Carolina at Chapel Hill.

Table of Contents

List of Tables and Figures

Controlled Vocabulary Use by Data Repositories:

Status and Potential for Promoting Interoperability

The DataNet Program funded by the National Science Foundation seeks to develop a sustainable infrastructure for data-driven research (National Science Foundation, 2007). Two complementary goals of this infrastructure are to promote discovery of data within and across existing repositories and to deter silo effects.

Controlled vocabularies are crucial for interoperability both within and across data management environments. Controlled vocabularies promote greater consistency and can contribute to an architecture supporting a unified set of services and interfaces. In service of these goals, the Helping Interdisciplinary Vocabulary Engineering (HIVE) approach supports the dynamic and interoperable application of controlled vocabularies.

This master's paper reports on the preliminary results of a Web survey developed in order to understand controlled vocabulary uses by data repository stakeholders and to identify how HIVE may better support stakeholder needs regarding controlled vocabularies.

1.1 Background

Controlled vocabularies continue to proliferate in connection with the growing data deluge (Willis, Greenberg, and White, 2012). Furthermore, data repositories face challenges related to using controlled vocabularies related to cost, interoperability, usability, and interdisciplinarity (Greenberg, Losee, Pérez Agüera, Scherle, White, and Willis, 2011). These challenges are magnified considerably when considered across data repositories rather than within a single data repository, as in cyberinfrastructure building efforts. It would be prohibitively expensive to attempt to maintain a nationally or internationally endorsed metadata vocabulary at the level of an NSF DataNet Partner.

The HIVE project aims to meet some of these challenges by providing an approach for integrating multiple controlled vocabularies and automatically generating metadata. A HIVE instance is populated with controlled vocabularies relevant to a data repository's community. Data contributors or data curators may then select terms from multiple controlled vocabularies in order to describe an item (e.g. a dataset or abstract or journal article). Terms may be selected by one of two HIVE components, either manually by means of a concept browser or automatically by means of an algorithm that suggests a set of candidate terms. After terms are selected, the item is indexed with those terms. Because its term-suggesting algorithm relies upon matching terms within an item to terms in the controlled vocabularies populating a HIVE instance, the HIVE approach is particularly well-suited for interdisciplinary data collections where textual components can be leveraged to aid suggestion of candidate terms across multiple controlled vocabularies.

Several large-scale stakeholder surveys funded by DataONE, one of the NSF-funded DataNet Partners, have examined attitudes toward research data services within particular groups of data repository stakeholders. Tenopir et al (2011) examined the attitudes and preferences of scientists toward data sharing. Subsequently Tenopir, Sandusky, Allard, and Birch (2013) examined attitudes of academic librarians toward research data services. However, little is known about controlled vocabulary uses across a broad swathe of data repository stakeholders e.g. data contributors, data curators, DataNet administrators, and repository developers. This master's paper seeks to make a contribution toward that research need.

1.2 Purpose

The purpose of this study was to describe controlled vocabulary uses of data repository stakeholders – data contributors, data curators, DataNet administrators, and repository developers – in order to better understand how to promote interoperability both within and among data repositories. Another significant purpose was the development of a framework for researching controlled vocabulary challenges and broader interoperability questions for data management. Greater insight into different stakeholders' uses of controlled vocabularies would enable the HIVE team to identify priorities for development and, ultimately, to provide more relevant controlled vocabulary services.

1.3 Research Questions

1. What controlled vocabularies are being used to describe research data?
2. What demand exists for HIVE-like services among data repository stakeholders?

2. Method

2.1 Research Design

The University of North Carolina at Chapel Hill Office of Human Research Ethics approved this study as a Web survey with the anonymity of participants protected. The survey was implemented using Qualtrics software. Findings are reported in the aggregate, and identifiers are stored separately from the survey data. Five pilot testers provided feedback on a first draft of the survey instrument, which was revised before dissemination. The survey was open for responses from May 17, 2013 to July 15, 2013; this master's paper reports preliminary analysis of responses collected through June 19, 2013. A convenience sample was used. An email survey invitation containing a survey link was distributed to project champions within each DataNet community as well as the following listservs associated with research data management:

- ACRL Digital Curation IG
- ACRL STS-L
- ALCTS Metadata IG
- CODATA
- DARTG
- DC-SAM
- EPA
- iPlant
- JE
- JISC Research Data Management

- LTER
- PAMWG
- RDA
- RDAP
- SE
- SIG-CR
- SIG-STI
- Taxonomic Data Working Group
- UNC Data Management WG
- USGS

Recipients were encouraged to forward the email survey invitation to relevant communities.

Given the email distribution method, there is no way of knowing the total number of survey link recipients. Ultimately, 180 recipients answered at least one question beyond Q2, the question which determined the path of the survey and which enabled many participants to determine whether they were a member of the target population. It is not unreasonable to estimate that the survey instrument reached 2,000 people, in which case the response rate would be approximately 9%. However, this study makes no claims to generalizability and instead aims to develop an improved survey instrument in order to study controlled vocabulary use across multiple roles associated with data repositories.

2.2 Survey Instrument

The survey instrument was designed with a branching question path, which consisted of the following seven sections:

- Consent to Participate in a Research Study

- Determining Survey Path
- Questions for Data Contributors
- Questions for Data Curators and Repository Developers
- Questions for Data Curators, DataNet Administrators, and Repository Developers
- Demographic Questions
- Concluding Questions

Participants' roles associated with data repositories determined the question path within the survey (see Figure 1. Survey Question Path). An early question in the survey, Q2, determined whether a participant identified as a data contributor, data curator, repository developer, and/or DataNet administrator; in recognition that participants may identify with multiple roles, the survey asked participants to select all that apply. Based upon his or her response, a participant would then progress to the blocks of questions associated with the roles with which the participant identified. If a participant identified with more than one role, that participant would be shown more than one block of questions. All participants were shown the demographic questions and concluding questions (an opportunity to provide feedback about the survey instrument or additional perspectives not specifically elicited by the survey). A complete version of the survey instrument including the logic determining which participants were shown which questions can be found in Appendix D: Survey Instrument.

In addition to role associated with data repository, other demographic questions included DataNet affiliation, length of involvement with a DataNet (if any), DataONE member node affiliation, and DFC project partner affiliation. Data contributors were asked to indicate primary and secondary fields of study as well as any DataONE member nodes or DFC data grids with which they had deposited data.

Within the block of questions shown to participants who identified as data contributors, participants were asked the following:

- from which controlled vocabularies they had selected terms when describing data deposited with any repository
- how frequently they had selected terms from a controlled vocabulary when describing data deposited with any repository
- which actions related to selecting terms from a controlled vocabulary they had performed
- which actions related to selecting terms from a controlled vocabulary they would perform in the next 12 months if that function were supported by the repository in which they were depositing data

The latter two questions were asked in order to gauge demand for the kind of controlled vocabulary services that HIVE can provide. In addition, a short series of questions were asked of data contributors in order to gauge attitudes regarding who (information professionals or data contributors) should provide terms describing research data and why.

The question block intended for data curators and repository developers closely resembled the question block intended for data contributors except that questions were framed in terms of the participant’s repository. Within this question block, participants were asked the following:

- which controlled vocabularies their repository uses
- which actions related to selecting terms from a controlled vocabulary their repository supports
- which actions related to selecting terms from a controlled vocabulary their repository would support in the next 12 months if it were possible

to support those actions now

- whether their repository uses any controlled vocabularies whose terms are represented as Uniform Resource Identifiers (URIs)
- whether their repository uses any controlled vocabularies whose terms are not represented as URIs
- whether their repository performs validation of terms selected by data contributors or data curators from controlled vocabularies

As with data contributors, some of these questions were intended to gauge demand for the kind of controlled vocabulary services HIVE can provide.

The final branch of the question path was intended for data curators, DataNet administrators, or repository developers – in other words, all repository staff who may be involved in the decision-making process to support certain controlled vocabulary services. This question block included one close-ended question asking participants to rate how eight aspects facilitate or impede their use of controlled vocabularies to describe scientific research data. In addition, two open-ended questions were asked:

- Has participation in a DataNet or other data repository influenced your plans for using controlled vocabularies? How?
- If a tool were to be built that would support the use of controlled vocabularies within and across DataNet Partners, what features would it need? How would you use such a tool?

This question block was designed in order to discover additional services that HIVE might provide in order to facilitate the use of controlled vocabularies within and across repositories.

2.3 Data Analysis

Data cleanup and analysis were performed using IBM SPSS Statistics software.

Data cleanup involved deleting all responses that did not answer a question beyond Q2 as these responses were likely participants who decided not to respond further after viewing the first questions in a role-specific question block. In addition, certain demographic variables were cleaned. For example, if a participant did not affiliate with a DataNet but did affiliate with a DataONE member node or a DFC project partner, the DataNet affiliation variable was cleaned using SPSS command syntax language.

Data analysis involved calculating descriptive statistics – primarily frequency counts. Two crosstabulations were performed in order to gauge demand for HIVE-like services. One crosstabulation compared which controlled vocabulary actions data contributors had performed in the past versus which controlled vocabulary actions data contributors would like to perform in the next twelve months. Another crosstabulation compared which controlled vocabulary actions data repositories currently support versus which controlled vocabulary actions data repositories would support in the next 12 months if it were possible to support those actions now. Finally, the means of variables related to facilitating and impeding controlled vocabulary use were calculated in order to determine which variables were facilitators and which were inhibitors. Responses to open-ended questions were not coded but were reviewed.

3. Results

This results section begins by providing an overview of respondent characteristics. It then provides a more detailed look at controlled vocabulary use across different roles associated with data repositories.

3.1 Demographics of Respondents

Many respondents identified with more than one role (see Table 1. Role Associated with Data Repository). Sixty-two percent (62%) of respondents identified as a repository developer. Forty-six percent (46%) of respondents identified as a data curator. Thirty percent (30%) of respondents identified as a data contributor. Twenty-five percent (25%) identified as a DataNet Administrator. Out of a total of 180 respondents, 329 answer choices associated with a role were selected, meaning that on average a respondent identified with 1.8 roles.

Of respondents affiliated with a DataNet, most were affiliated with DataONE (17.8%) or the DFC (13.9%); however, most respondents (66.1%) had no DataNet affiliation (see Table 2). Every DataONE member node was represented (see Table 4); almost every DFC project partner was represented (see Table 3).

Respondents who were data contributors represent a variety of science and social science disciplines, with most from information and library science. The breakdown of respondents by disciplines is provided in Table 5. Respondents who were data contributors had deposited data with six out of 11 DataONE member nodes (see Table 7) and three out of six DFC data grids (see Table 6).

Overall, the demographics of respondents demonstrate that this study’s sample is not representative of the broader population of data repository stakeholders. Nevertheless, this convenience sample was sufficient to indicate ways in which the survey instrument could be improved and suggest potential hypotheses if the study were to be revised and conducted on a larger scale.

3.2 Data Contributors

Data contributors were asked from which controlled vocabularies they had selected terms when describing data deposited with any repository by means of both a closed-ended question and an open-ended question. In response to the closed-ended question, the vocabularies most used were LCSH, NBII, EnvThes and/or LTER, ITIS, MeSH, and TGN (see Table 8). However, responses to the open-ended question indicated a much longer tail of vocabularies in use. A total of 25 additional vocabularies were supplied by data contributors, of which 18 were named by only one respondent.

Another interesting aspect of the responses to the open-ended question were how many “vocabularies” supplied by participants were not vocabularies at all but rather a metadata standard identifying fields and relationships among fields, e.g. Dublin Core and Darwin Core. Eliminating non-vocabularies from the responses would require careful analysis of those less familiar to the researcher, e.g. the SPASE (meta)Data Model or the W3C Provenance Ontology (PROV-O).

A crosstabulation was performed to compare which controlled vocabulary actions data contributors had performed in the past versus which controlled vocabulary actions data contributors would like to perform in the future (see Table 10). Of 19 data contributors who had not in the past selected from multiple controlled vocabularies when describing a single dataset, 14 indicated that they would do so in the next 12 months. Of 30 data contributors who had not in the past used software to generate suggested terms selected from a controlled vocabulary, 22 indicated that they would do so in the next 12 months. Although the sample of this study is not representative, these results suggest that a future study might hypothesize a demand for HIVE-like services.

3.3 Data Curators and Repository Developers

Just as data contributors were asked from what controlled vocabularies they had selected terms when depositing data, data curators and repository developers were asked what controlled vocabularies their repositories use by means of both a closed-ended question and an open-ended question. In response to the closed-ended question, the vocabularies most used were the same as those indicated by data contributors, albeit in a different order – LCSH, MeSH, TGN, ITIS, NBII, and EnvThes and/or LTER (see Table 11). However, responses to the open-ended question indicated an even longer tail of vocabularies in use than that indicated by data contributors. An astonishing total of 60 additional vocabularies were supplied by data curators and repository developers, of which 44 were named by only one respondent. The top three vocabularies supplied by data curators and repository developers were the NASA GCMD Earth Science Keywords (frequency=13), the NetCDF Climate and Forecast (CF) Metadata Convention (frequency=11), and ISO 19115 Topic Categories (frequency=7). Notably, the NASA GCMD Earth Science Keywords also appeared in the top three vocabularies of those supplied by data contributors (see Table 9).

As with the vocabularies supplied by data contributors, eliminating non-vocabularies from those identified by data curators and repository developers would require careful analysis of those less familiar to the researcher. Explicit assumptions would have to be made about how to determine what constitutes a vocabulary and what does not. Furthermore, a repository might use terms derived from an ontology or a classification system as terms in a custom controlled vocabulary. For these reasons no vocabularies supplied by data contributors or data curators and repository developers were eliminated from this analysis.

A crosstabulation was performed to compare which controlled vocabulary actions repositories currently support versus which controlled vocabulary actions they would like to support in the future (see Table 13). Of 41 data curators and repository developers whose repository does not currently support selecting from multiple controlled vocabularies when describing a single dataset, 22 indicated that their repository would do so in the next 12 months if it were possible to support those actions now. Of 59 data curators and repository developers whose repository does not currently support using software to generate suggested terms selected from a controlled vocabulary, 28 indicated that their repository would do so in the next 12 months if it were possible to support those actions now. Within this sample, data curators and repository developers are noticeably more circumspect when indicating future support of a service than data contributors are when indicating future use of a service. Whether this reserve is due to reluctance to speak for one's repository versus oneself, a pragmatic view of organizational resources, differing attitudes about automatic metadata generation, or other variables remains unclear. Even so, these results suggest that a future study might hypothesize a demand for HIVE-like services among data curators and repository developers as well as data contributors.

Forty-one percent (41%) of respondents' data repositories make use of controlled vocabularies whose terms are represented as URIs, but a majority (53%) make use of controlled vocabularies whose terms are not represented as URIs (see Tables 14 and 15). Some overlap occurs, indicating some repositories that make use of vocabularies whose terms are represented as URIs as well as vocabularies whose terms are not. Thirty-seven percent (37%) of respondents' data repositories perform validation of terms selected to describe data deposited with their repository against specific controlled vocabularies (see Table 16).

3.4 Data Curators, DataNet Administrators, and Repository Developers

Facilitators and inhibitors of controlled vocabulary use among data curators, DataNet administrators, and repository developers were not conclusively identified. Participants were asked to rate how eight aspects facilitate or impede their use of controlled vocabularies to describe scientific research data on a five-point scale with 1 indicating “Greatly impede,” 3 indicating “Neither facilitate nor impede,” and 5 indicating “Greatly facilitate”. The eight aspects were as follows:

- Local or in-house governance of a controlled vocabulary
- National or international governance governance of a controlled vocabulary
- Availability of a controlled vocabulary on the World Wide Web
- Availability of a controlled vocabulary’s terms as URIs
- Data storage for a controlled vocabulary (e.g. spreadsheet, relational database, thesaurus software, Web)
- Currency or update frequency of a controlled vocabulary
- Openness of a controlled vocabulary’s governance to term suggestions
- Ability to generate suggested subject terms selected from a controlled vocabulary

With the exception of availability on the World Wide Web, which had a mean of 4.20, means of these aspects ranged from 3.27 to 3.88. With little variation among these values, which hover between “Neither facilitate nor impede” and “Somewhat facilitate” on the five-point scale, none of the eight aspects can be conclusively identified as either a facilitator or an inhibitor.

An open-ended question asked data curators, DataNet administrators, and repository developers “If a tool were to be built that would support the use of controlled vocabularies within and across DataNet Partners, what features would it need? How would you use such a tool?” (see Table 18). Qualitative coding of the responses was not performed; however, the responses were reviewed with an eye toward revising the survey instrument.

Even without qualitative coding, several themes emerge. One theme is the importance of web services:

- “An open, well-documented API that would allow validation against CVs. It would be nice if the validation source could be local, so we could have a local copy of the CV for fast validation. It would also be nice if CVs could be expressed in a standard format so that we could add custom CVs to our validation repository or adapt existing CVs thare [sic] are not in popular use among other DataNet partners. In other words, a pretty general API that automatically supports many popular CV standards, but also allows for custom/unpopular CVs to be used.”
- “Ease of use, ease of ‘plugging’ into different services and software.”
- “It would require availability of vocabulary’s terms as Uniform Resource Identifiers (URIs).”
- “We would be more likely to use the tool if it was offered in the form of a web services API as opposed to a website or a desktop application. Web services would make the tool platform-independent and easier to embed within our current suite of software application[s].”

A second theme is the ability to simultaneously manage controlled and uncontrolled vocabularies or internal and external vocabularies.

- "For our dataset it would require the abilty [sic] to manage our own terms in addition to [external] controlled vocabularies, as consistency with our primary data users is more important than adherence to a controlled vocabulary that doesn't meet all of our needs and/or isn't used by our data users."
- "Given user provided abstracts and keywords using an uncontrolled vocabulary, we need to parse the user generated input into a controlled structure. We would use the tool to populate search indices."
- "I would use such a tool to add preferred terms to records while keeping free-text tags in place."

A third theme is the ability to capture metadata earlier in the data lifecycle:

- "Generation of metadata from the workflows and applications that generate the data."
- "A DataNet tool needs to be something that easily expands beyond DataNets and which facilitates the use of existing vocabularies, particularly at the data generation stage."

A fourth theme is the ability to crosswalk or map between terms from different vocabularies:

- "Registries, ontology mapping, annotation. Would use it to map between concepts and to describe limitations of those mappings."
- "Some level of ontology mapping between overlapping vocabularies is necessary."

- “I need mappings between controlled vocabularies for different communities.”
- “Ideally, disambiguation between similar terms with different usages and differing terms with similar semantics.”

Taken together, these themes suggest potential aspects to investigate as facilitators of controlled vocabulary use if this study were to be revised and implemented on a larger scale.

4. Discussion

4.1 Conclusions

A companion effort to the ongoing development of the HIVE project, this research study gathered information about controlled vocabulary use across many different sets of data repository stakeholders – data contributors, data curators, DataNet administrators, and repository developers.

The first research question of this study asked what controlled vocabularies are being used to describe research data. Twenty-five (25) vocabularies were supplied by data contributors, of which 18 were named by only one respondent (seeTable 9). Sixty (60) additional vocabularies were supplied by data curators and repository developers, of which 44 were named by only one respondent (see Table 12). Across data contributors, data curators, and repository developers, the top three vocabularies supplied by respondents were the NASA GCMD Earth Science Keywords (frequency=13), the NetCDF Climate and Forecast (CF) Metadata Convention (frequency=11), and ISO 19115 Topic Categories (frequency=7) – all of which are located squarely within DataONE target disciplines. The long tail of controlled vocabularies actively in use by data repositories affirms the design decision of HIVE to allow each instance to import vocabularies selected for use by that repository's community.

The second research question of this study asked what demand exists for HIVE-like services among data repository stakeholders. Of 19 data contributors who had not in the past selected from multiple controlled vocabularies when describing a single dataset, 14 indicated that they would do so in the next 12 months. Of 30 data contributors who had not in the past used software to generate suggested terms selected from a controlled vocabulary, 22 indicated that they would do so in the next 12 months. Of 41 data curators and repository developers whose repository does not currently support selecting from multiple controlled vocabularies when describing a single dataset, 22 indicated that their repository would do so in the next 12 months if it were possible to support those actions now. Of 59 data curators and repository developers whose repository does not currently support using software to generate suggested terms selected from a controlled vocabulary, 28 indicated that their repository would do so in the next 12 months if it were possible to support those actions now. This study does not claim generalizability to the broader population of data repository stakeholders from its convenience sample. Even so, these results (see Table 10 and Table 13) suggest that a future study might hypothesize a demand for HIVE-like services among data curators and repository developers as well as data contributors.

Identifying facilitators and inhibitors of controlled vocabulary use is related to the question of what demand exists for HIVE-like services. However, facilitators and inhibitors of controlled vocabulary use were not conclusively identified.

Arguably the most important output of this research study was the development of a framework for studying controlled vocabulary use across different roles associated with data repositories. Two major revisions to the survey instrument are recommended in the event that the study is revised and implemented on a larger scale:

- Remove the “Other [Please specify]” option from Q2, responses to which determine the question path followed by a respondent. This design leaves open the possibility that someone who might select the answer choice associated with a defined role will instead select only the “Other [Please specify]” answer choice. If this happens, a respondent is not shown any of the question blocks associated with a repository role. If the researchers wish to understand more specifically how respondents characterize their role, a subsequent open-ended question could be added.

- Redesign Q22, which asks participants to rate how eight aspects facilitate or impede their use of controlled vocabularies to describe research data. The aspects enumerated by the question could be revised keeping in mind responses to the open-ended question “If a tool were to be built that would support the use of controlled vocabularies within and across DataNet Partners, what features would it need? How would you use such a tool?” (see Table 18). Additionally, each aspect should be parsed into two opposite aspects. For example, the aspect “Currency or update frequency” could be parsed into “Frequent updates” and “Infrequent updates,” each of which participants would rate on a five-point scale. In this way, responses to “Frequent updates” could validate responses to “Infrequent updates” and vice versa.

4.2 Limitations and Future Research

The primary limitation of this research study is its convenience sample, which prevents the study from being able to claim generalizability to the broader population of data repository stakeholders.

However, the study does reveal rich avenues for future research. With a revised survey instrument and more purposeful sampling, this study could produce a list of controlled vocabularies in use in the broader population of data repository stakeholders.

This list could be analyzed to determine which vocabularies adhere to which vocabulary development standards, which vocabularies have been encoded in SKOS, what work would need to be done in order for each vocabulary to be imported into a HIVE instance, and which vocabularies are the highest priority for the greatest swathe of stakeholders.

Interestingly, responses to both the open-ended questions asking respondents to identify controlled vocabularies in use or desired features of a vocabulary tool suggest the need to analyze vocabularies for describing data collection or data analysis – e.g. instrument lists, parameters, and micro-services – in addition to vocabularies for describing the subject of a dataset.

Analyzing vocabularies in use by data repository stakeholders could enable NSF DataNet Partners and other data repository stakeholders to more deeply understand the status and potential of controlled vocabularies for promoting interoperability among data repositories.

References

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.

Greenberg, J., Losee, R., Pérez Agüera, J.R., Scherle, R., White, H., and Willis, C. (2011). HIVE: Helping Interdisciplinary Vocabulary Engineering. *Bulletin of the American Society for Information Science and Technology, 37*(4). Retrieved from http://www.asis.org/BulletinApr-11AprMay11_Greenberg_etAl.html

Helping Interdisciplinary Vocabulary Engineering (HIVE) Demonstration System. Retrieved from http://hive.nescent.org/

Helping Interdisciplinary Vocabulary Engineering (HIVE) Wiki. Retrieved from https://www.nescent.org/sites/hive/Main_Page

National Science Foundation, Office of Cyberinfrastructure Directorate for Computer & Information Science & Engineering. (2007). Sustainable digital data preservation and access network partners (DataNet) program summary. Retrieved from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141

Tenopir, C. Allard, S., Douglass, K., Aydinoglu A.U., Wu L., et al. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE, 6*(6): e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Sandusky, R.J., Allard, S., and Birch, B. (2013). Academic librarians and research data services: Preparation and attitudes. *IFLA Journal, 39*(1): 70–78.

Willis, C., Greenberg, J., and White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology, 63*(8): 1505–1520.

Appendix A: Tables and Figures

Figure 1. Survey Question Path

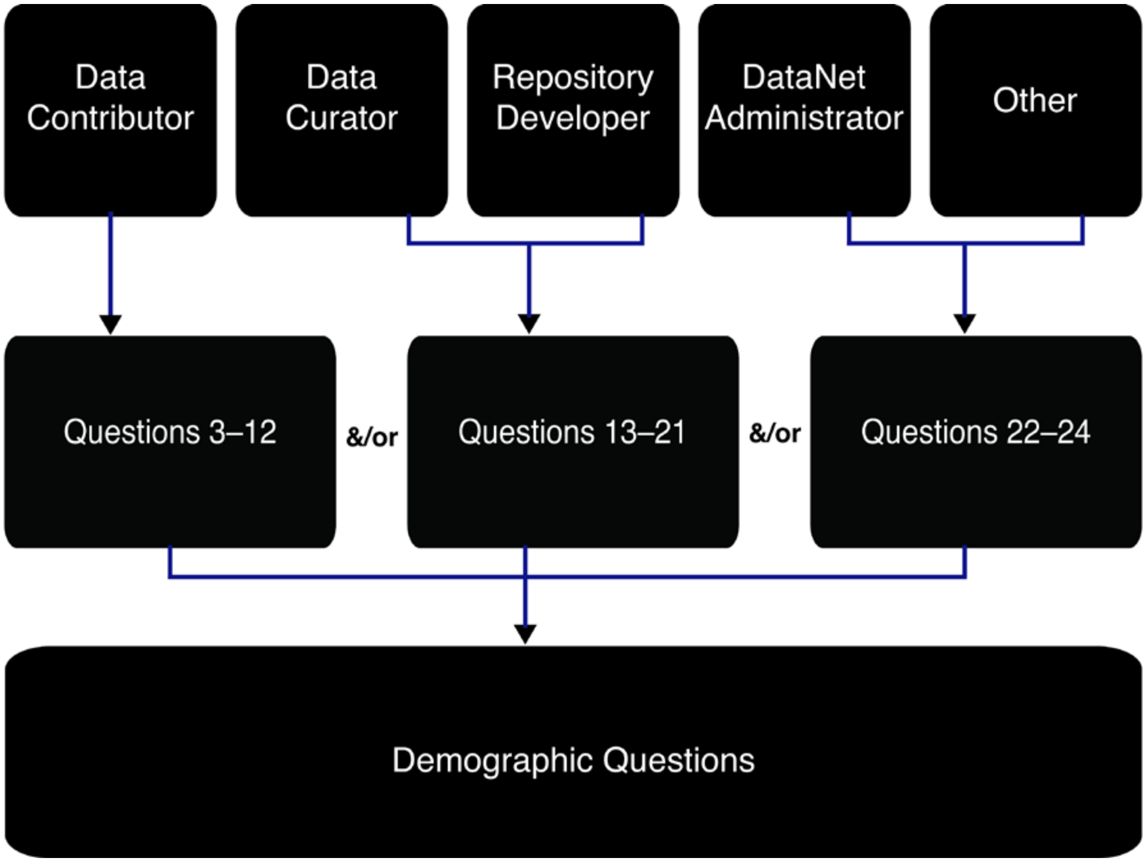
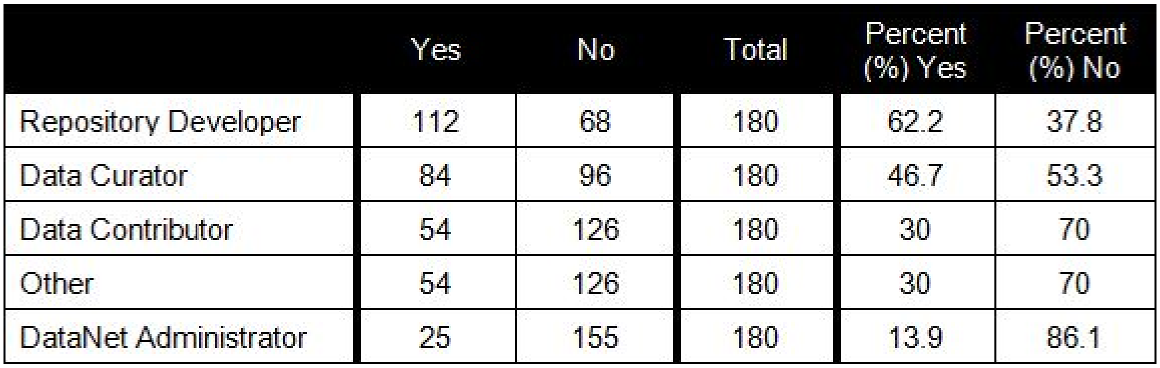


Table 1. Role Associated with Data Repository

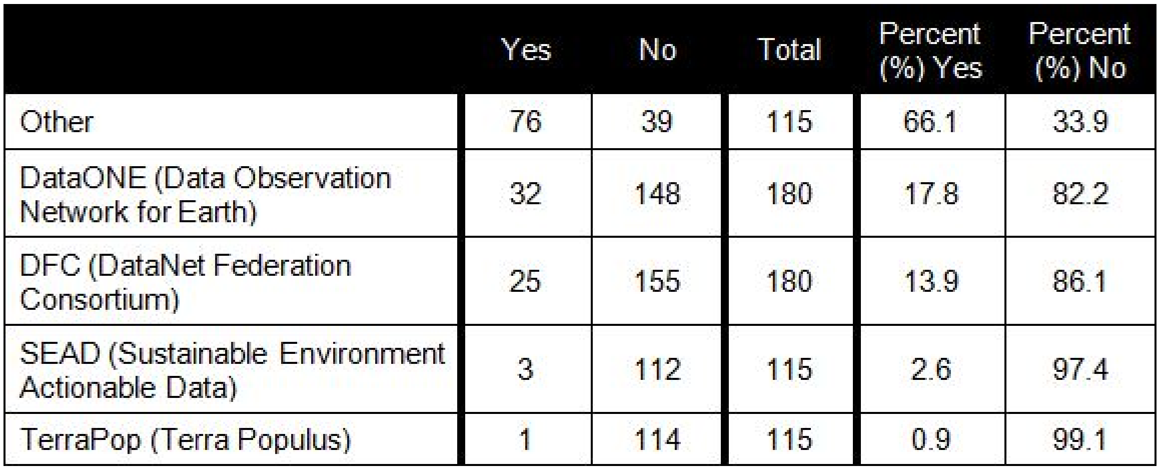
Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Repository Developer	112	68	180	62.2	37.8
Data Curator	84	96	180	46.7	53.3
Data Contributor	54	126	180	30	70
Other	54	126	180	30	70
DataNet Administrator	25	155	180	13.9	86.1

Table 2. NSF DataNet Partner Affiliation of Data Curators, Repository Developers, and DataNet Administrators

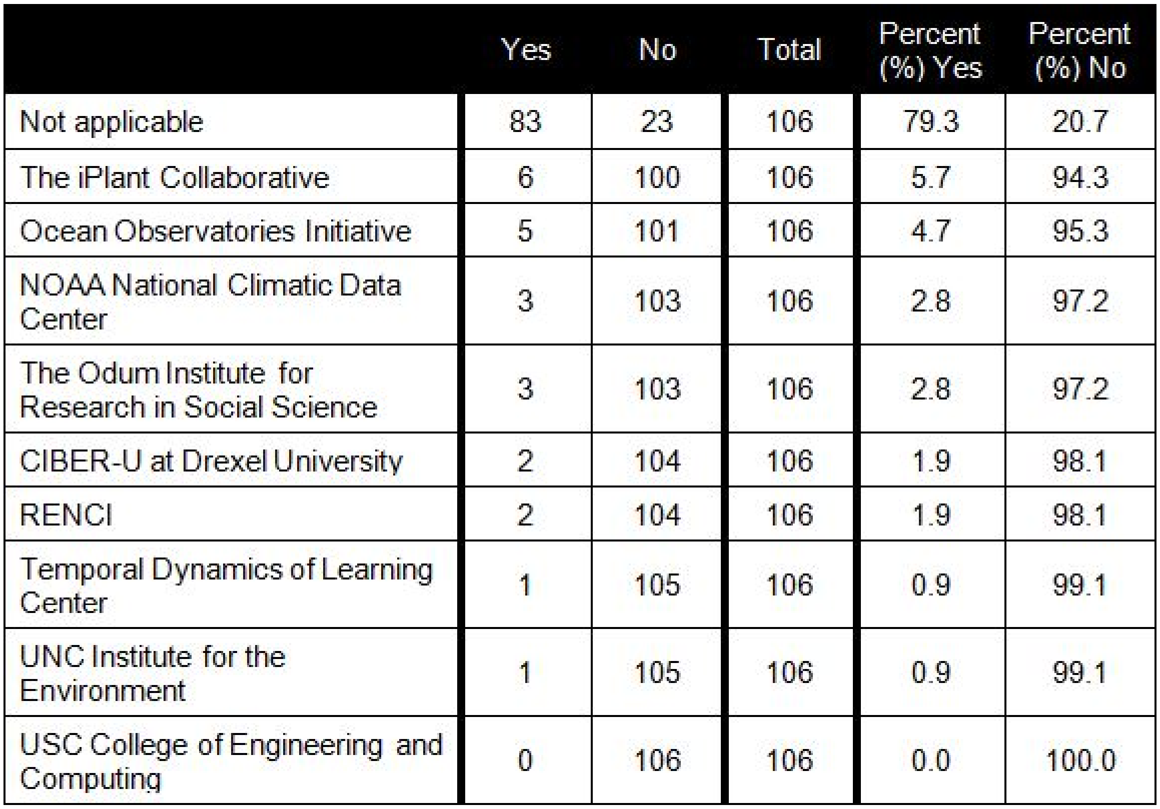
Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Other	76	39	115	66.1	33.9
DataONE (Data Observation Network for Earth)	32	148	180	17.8	82.2
DFC (DataNet Federation Consortium)	25	155	180	13.9	86.1
SEAD (Sustainable Environment Actionable Data)	3	112	115	2.6	97.4
TerraPop (Terra Populus)	1	114	115	0.9	99.1

Table 3. DFC Project Partner Affiliation of Data Curators, Repository Developers, and DataNet Administrators

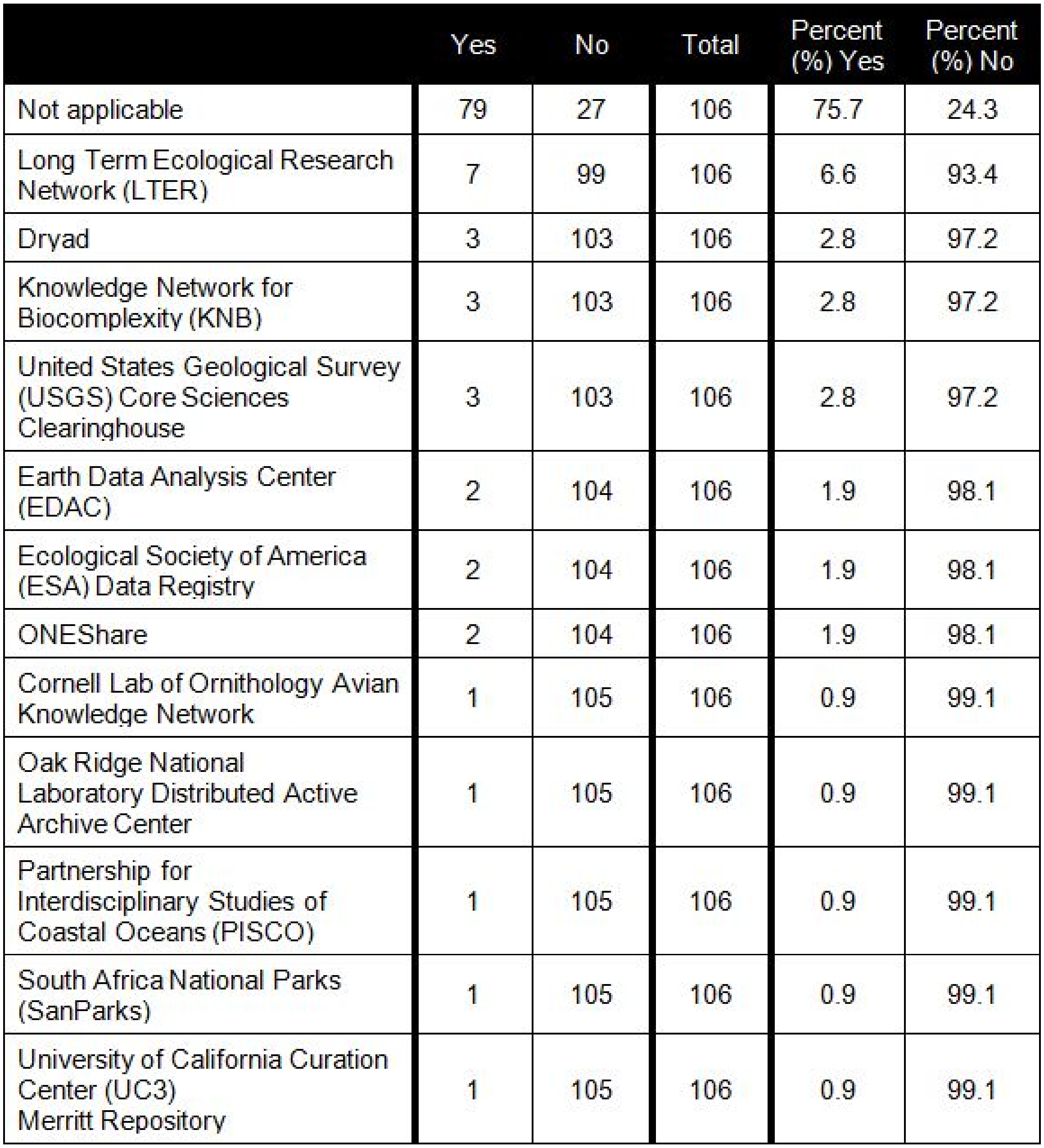
Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Not applicable	83	23	106	79.3	20.7
The iPlant Collaborative	6	100	106	5.7	94.3
Ocean Observatories Initiative	5	101	106	4.7	95.3
NOAA National Climatic Data Center	3	103	106	2.8	97.2
The Odum Institute for Research in Social Science	3	103	106	2.8	97.2
CIBER-U at Drexel University	2	104	106	1.9	98.1
RENCI	2	104	106	1.9	98.1
Temporal Dynamics of Learning Center	1	105	106	0.9	99.1
UNC Institute for the Environment	1	105	106	0.9	99.1
USC College of Engineering and Computing	0	106	106	0.0	100.0

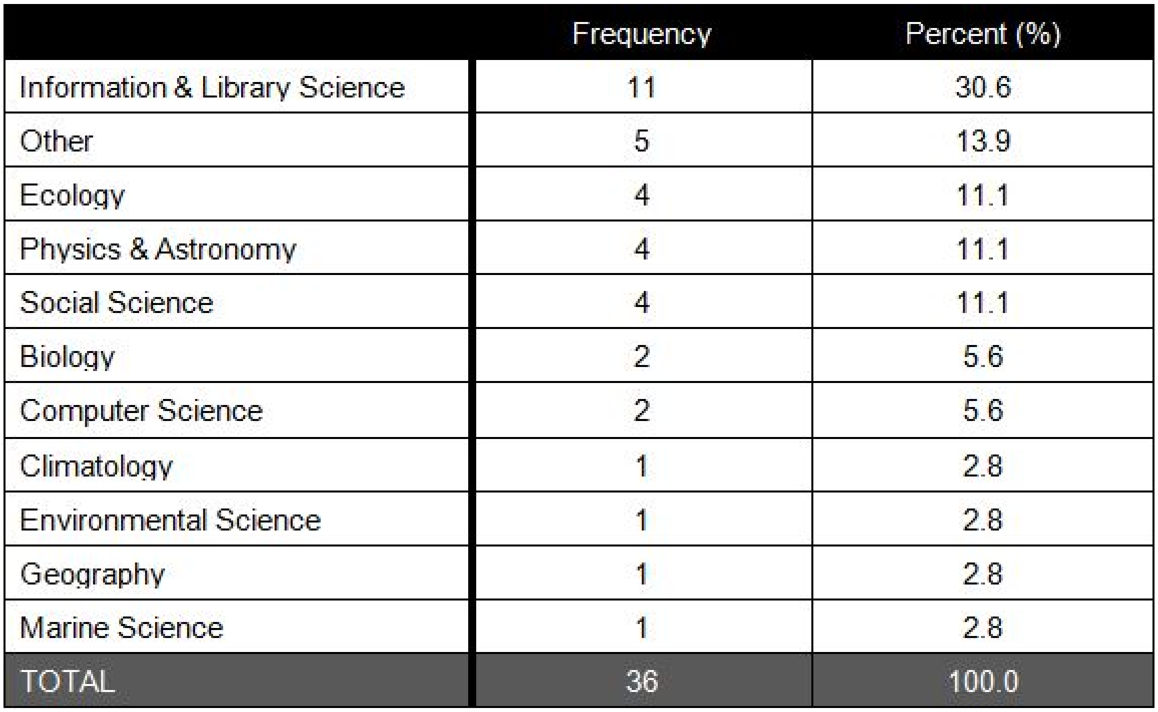
Table 4. DataONE Member Node Affiliation of Data Curators, Repository Developers, and DataNet Administrators

Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Not applicable	79	27	106	75.7	24.3
Long Term Ecological Research Network (LTER)	7	99	106	6.6	93.4
Dryad	3	103	106	2.8	97.2
Knowledge Network for Biocomplexity (KNB)	3	103	106	2.8	97.2
United States Geological Survey (USGS) Core Sciences Clearinghouse	3	103	106	2.8	97.2
Earth Data Analysis Center (EDAC)	2	104	106	1.9	98.1
Ecological Society of America (ESA) Data Registry	2	104	106	1.9	98.1
ONEShare	2	104	106	1.9	98.1
Cornell Lab of Ornithology Avian Knowledge Network	1	105	106	0.9	99.1
Oak Ridge National Laboratory Distributed Active Archive Center	1	105	106	0.9	99.1
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)	1	105	106	0.9	99.1
South Africa National Parks (SanParks)	1	105	106	0.9	99.1
University of California Curation Center (UC3) Merritt Repository	1	105	106	0.9	99.1

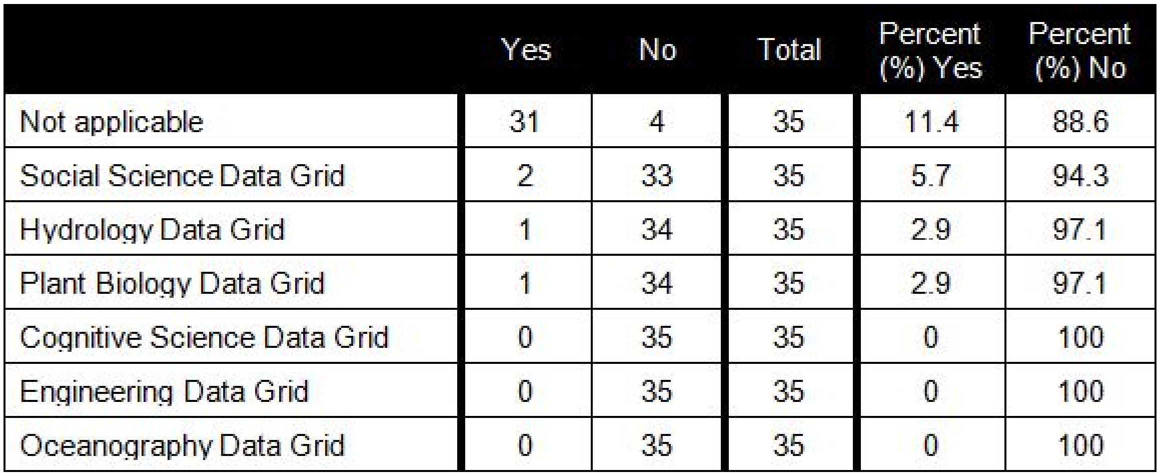
Table 5. Field of Study of Data Contributors



	Frequency	Percent (%)
Information & Library Science	11	30.6
Other	5	13.9
Ecology	4	11.1
Physics & Astronomy	4	11.1
Social Science	4	11.1
Biology	2	5.6
Computer Science	2	5.6
Climatology	1	2.8
Environmental Science	1	2.8
Geography	1	2.8
Marine Science	1	2.8
TOTAL	36	100.0

Table 6. DFC Data Grids in which Data Contributors Have Deposited Data

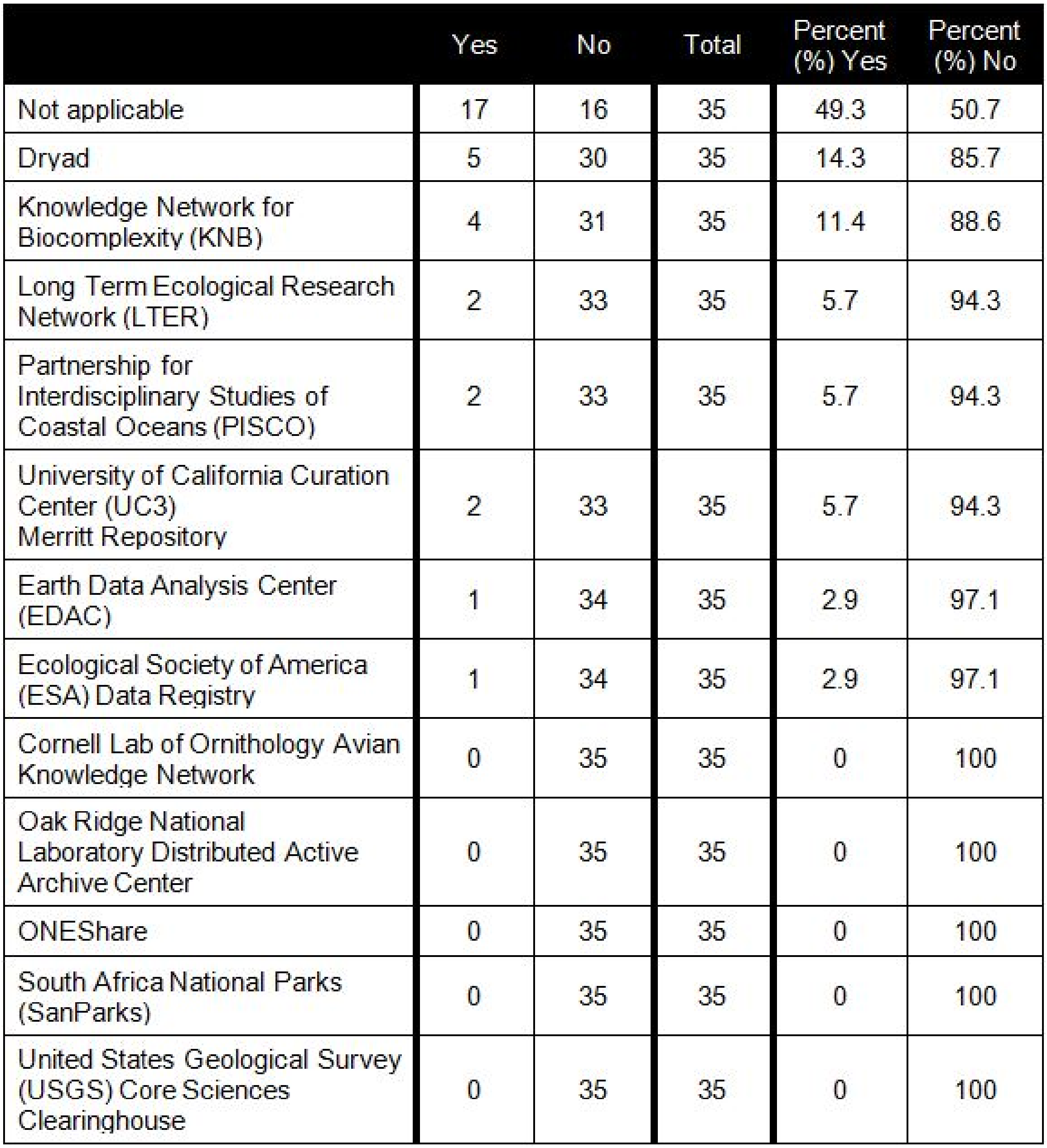
Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Not applicable	31	4	35	11.4	88.6
Social Science Data Grid	2	33	35	5.7	94.3
Hydrology Data Grid	1	34	35	2.9	97.1
Plant Biology Data Grid	1	34	35	2.9	97.1
Cognitive Science Data Grid	0	35	35	0	100
Engineering Data Grid	0	35	35	0	100
Oceanography Data Grid	0	35	35	0	100

Table 7. DataONE Member Nodes in which Data Contributors Have Deposited Data

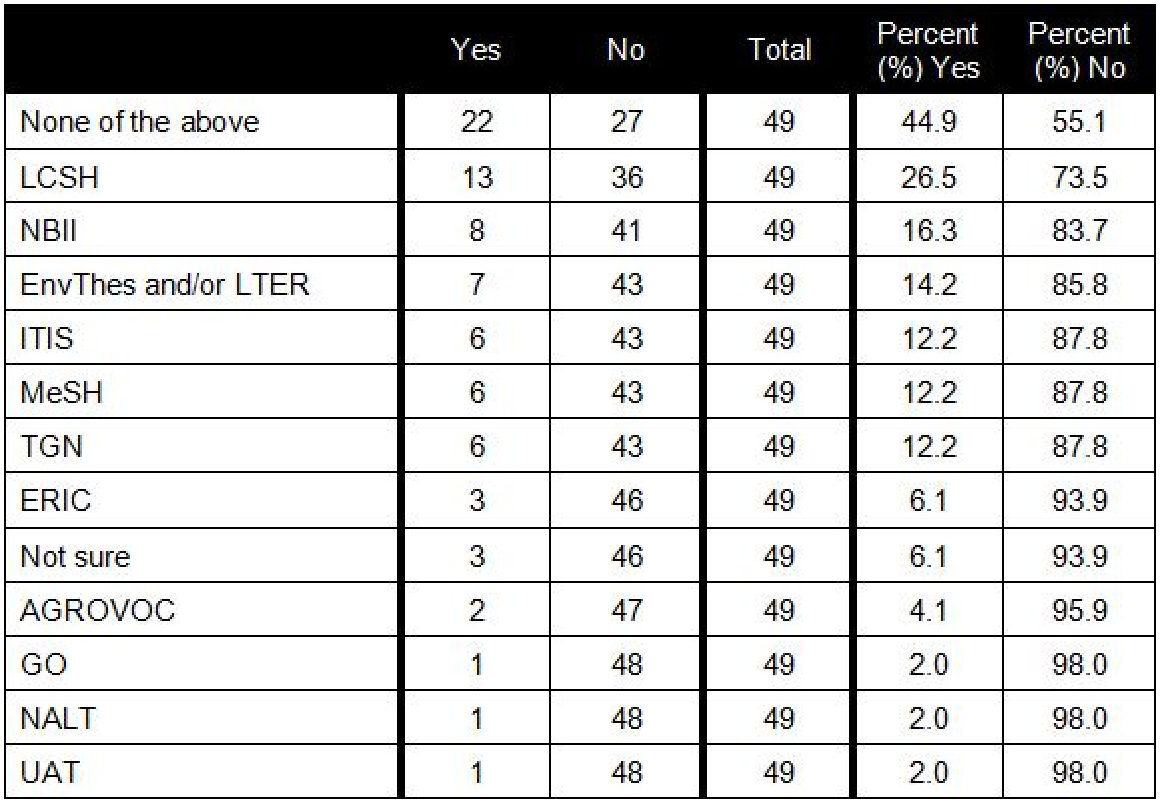
Participants were asked to select all that apply.



	Yes	No	Total	Percent (%) Yes	Percent (%) No
Not applicable	17	16	35	49.3	50.7
Dryad	5	30	35	14.3	85.7
Knowledge Network for Biocomplexity (KNB)	4	31	35	11.4	88.6
Long Term Ecological Research Network (LTER)	2	33	35	5.7	94.3
Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)	2	33	35	5.7	94.3
University of California Curation Center (UC3) Merritt Repository	2	33	35	5.7	94.3
Earth Data Analysis Center (EDAC)	1	34	35	2.9	97.1
Ecological Society of America (ESA) Data Registry	1	34	35	2.9	97.1
Cornell Lab of Ornithology Avian Knowledge Network	0	35	35	0	100
Oak Ridge National Laboratory Distributed Active Archive Center	0	35	35	0	100
ONEShare	0	35	35	0	100
South Africa National Parks (SanParks)	0	35	35	0	100
United States Geological Survey (USGS) Core Sciences Clearinghouse	0	35	35	0	100

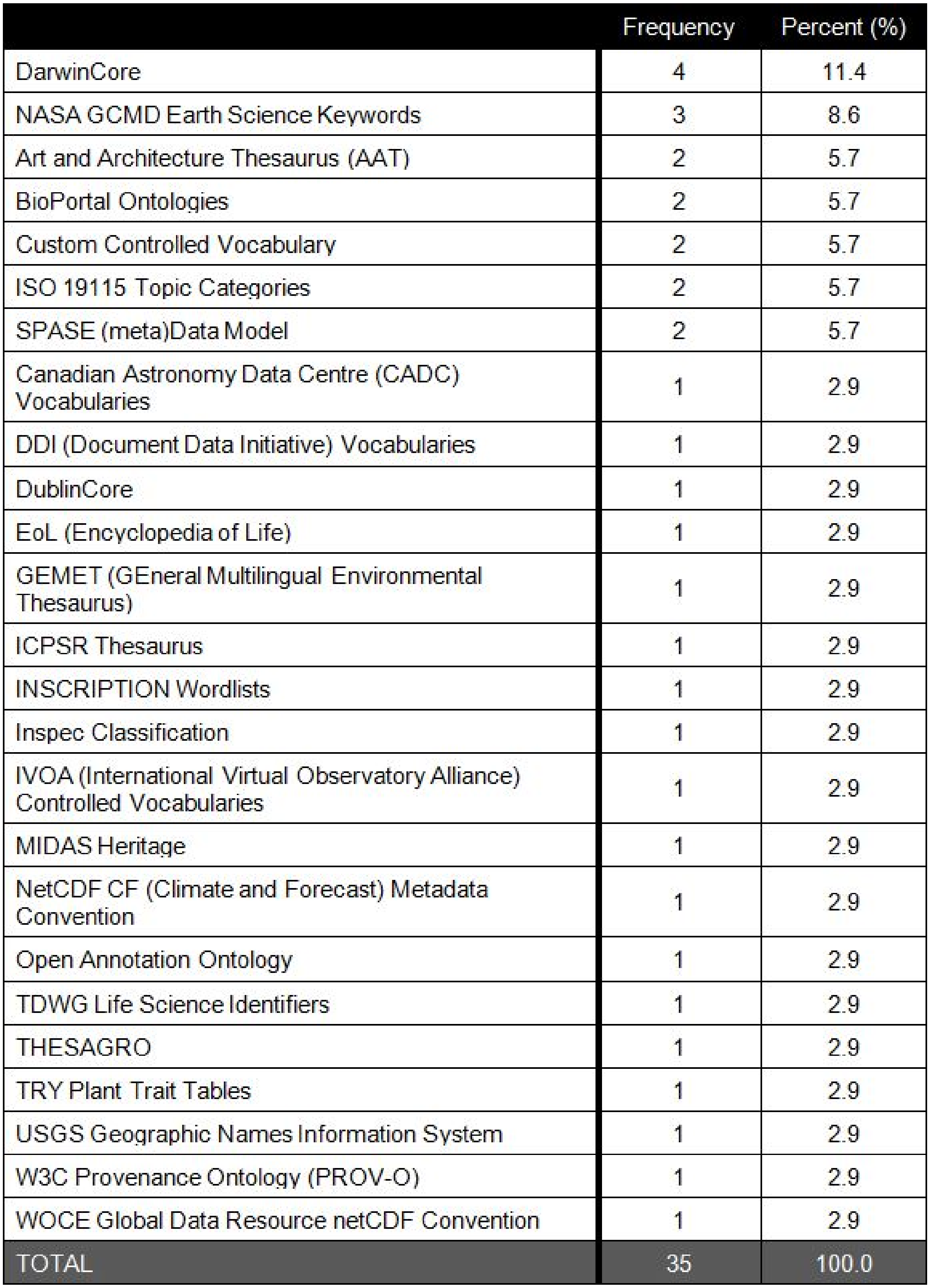
Table 8. Controlled Vocabularies Used by Data Contributors: Choices Supplied by Survey

Participants were asked to select all that apply.



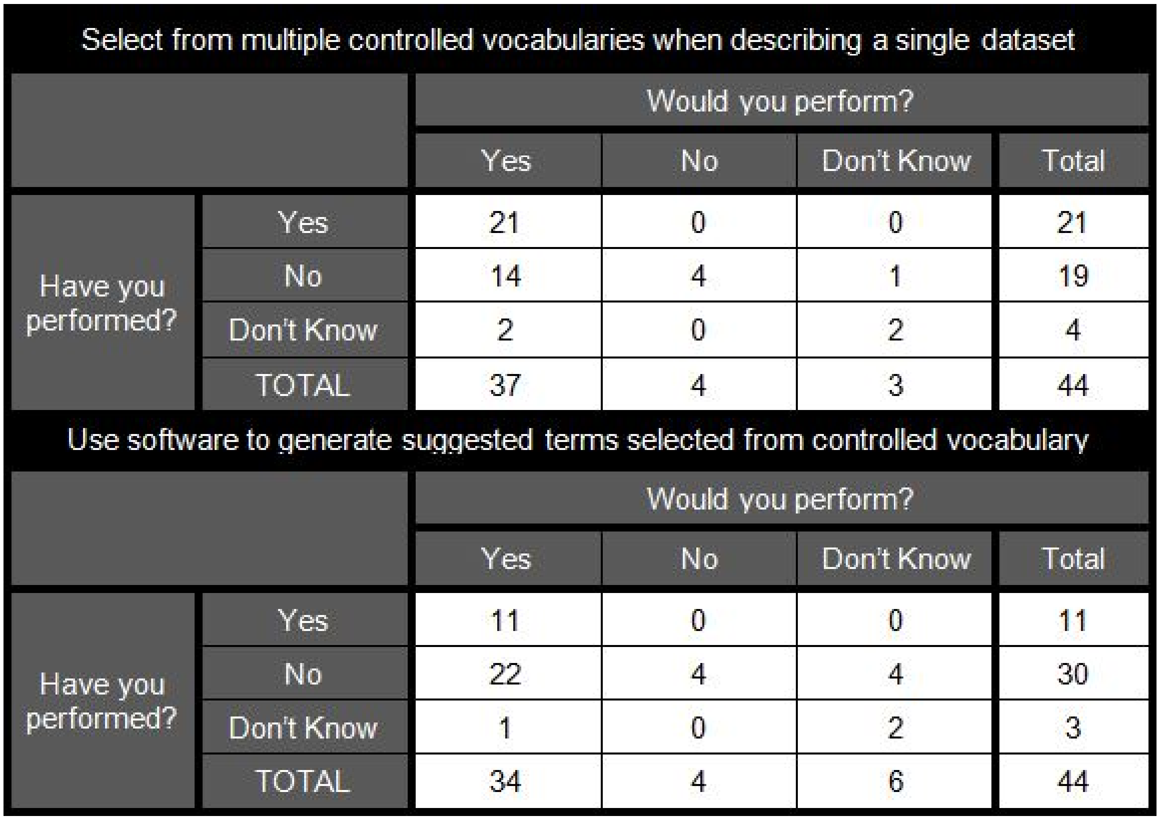
	Yes	No	Total	Percent (%) Yes	Percent (%) No
None of the above	22	27	49	44.9	55.1
LCSH	13	36	49	26.5	73.5
NBII	8	41	49	16.3	83.7
EnvThes and/or LTER	7	43	49	14.2	85.8
ITIS	6	43	49	12.2	87.8
MeSH	6	43	49	12.2	87.8
TGN	6	43	49	12.2	87.8
ERIC	3	46	49	6.1	93.9
Not sure	3	46	49	6.1	93.9
AGROVOC	2	47	49	4.1	95.9
GO	1	48	49	2.0	98.0
NALT	1	48	49	2.0	98.0
UAT	1	48	49	2.0	98.0

Table 9. Controlled Vocabularies Used by Data Contributors: Answers Supplied by Participants



	Frequency	Percent (%)
DarwinCore	4	11.4
NASA GCMD Earth Science Keywords	3	8.6
Art and Architecture Thesaurus (AAT)	2	5.7
BioPortal Ontologies	2	5.7
Custom Controlled Vocabulary	2	5.7
ISO 19115 Topic Categories	2	5.7
SPASE (meta)Data Model	2	5.7
Canadian Astronomy Data Centre (CADC) Vocabularies	1	2.9
DDI (Document Data Initiative) Vocabularies	1	2.9
DublinCore	1	2.9
EoL (Encyclopedia of Life)	1	2.9
GEMET (GEneral Multilingual Environmental Thesaurus)	1	2.9
ICPSR Thesaurus	1	2.9
INSCRIPTION Wordlists	1	2.9
Inspec Classification	1	2.9
IVOA (International Virtual Observatory Alliance) Controlled Vocabularies	1	2.9
MIDAS Heritage	1	2.9
NetCDF CF (Climate and Forecast) Metadata Convention	1	2.9
Open Annotation Ontology	1	2.9
TDWG Life Science Identifiers	1	2.9
THESAGRO	1	2.9
TRY Plant Trait Tables	1	2.9
USGS Geographic Names Information System	1	2.9
W3C Provenance Ontology (PROV-O)	1	2.9
WOCE Global Data Resource netCDF Convention	1	2.9
TOTAL	35	100.0

Table 10. Crosstabulation of Controlled Vocabulary Actions Performed by Data Contributors



Select from multiple controlled vocabularies when describing a single dataset

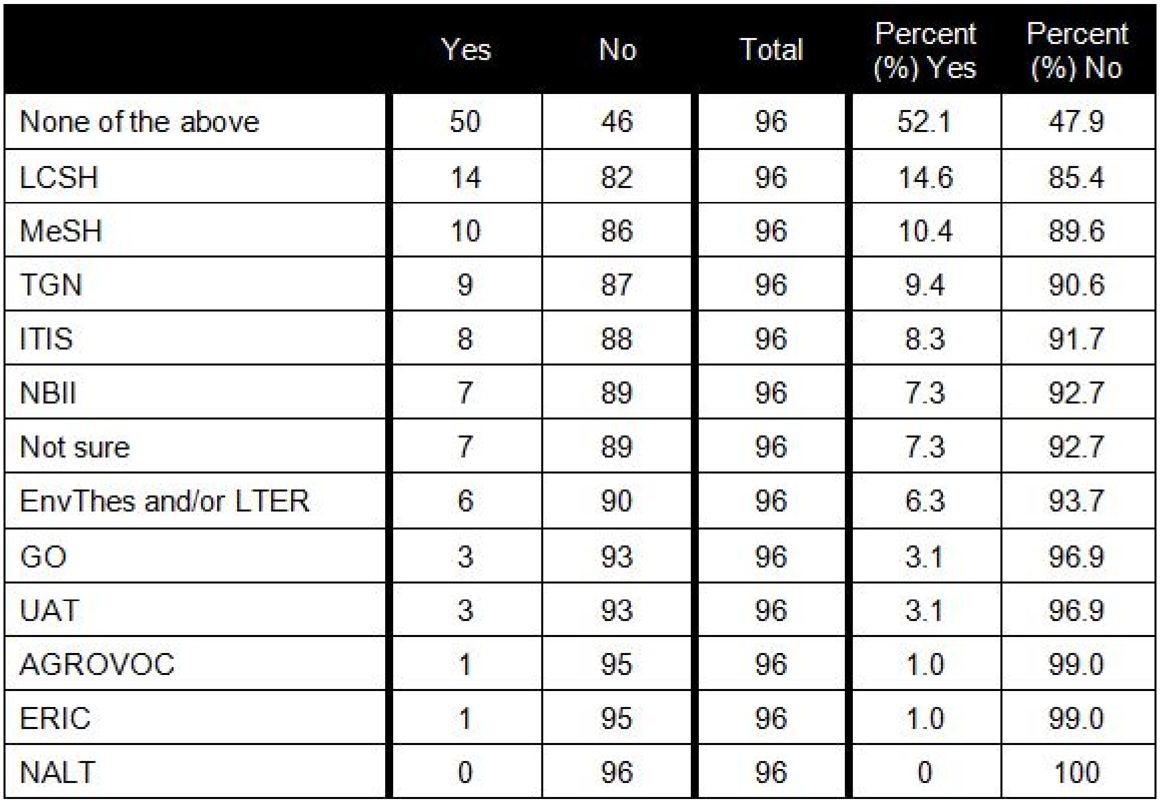
		Would you perform?			
		Yes	No	Don't Know	Total
Have you performed?	Yes	21	0	0	21
	No	14	4	1	19
	Don't Know	2	0	2	4
	TOTAL	37	4	3	44

Use software to generate suggested terms selected from controlled vocabulary

		Would you perform?			
		Yes	No	Don't Know	Total
Have you performed?	Yes	11	0	0	11
	No	22	4	4	30
	Don't Know	1	0	2	3
	TOTAL	34	4	6	44

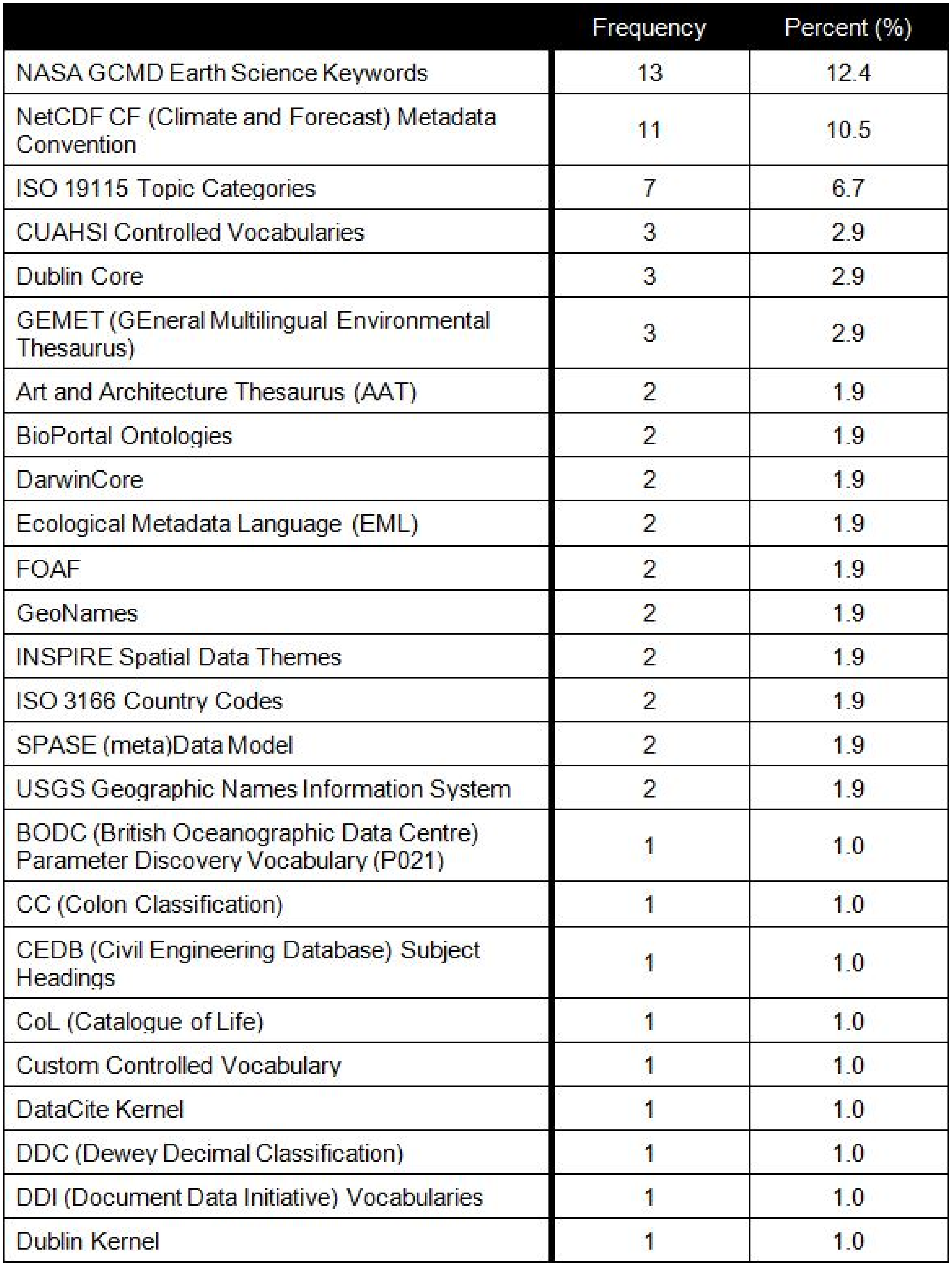
Table 11. Controlled Vocabularies Used by Data Curators and Repository Developers: Choices Supplied by Survey

Participants were asked to select all that apply.



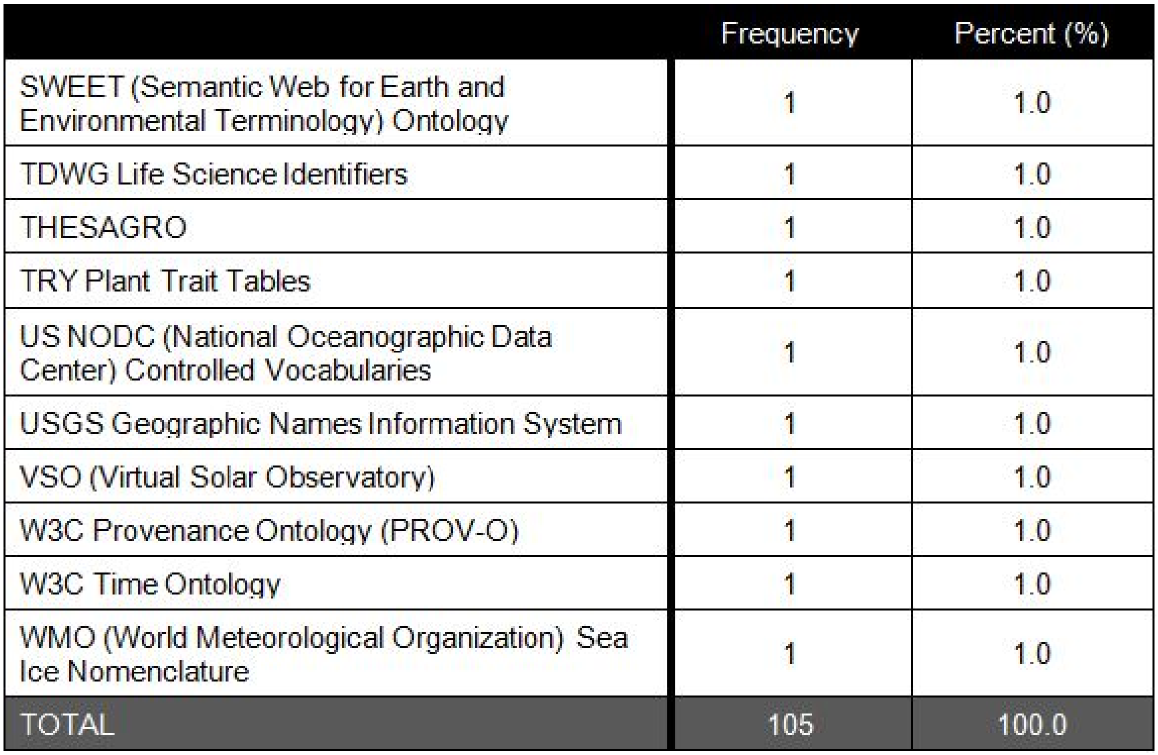
	Yes	No	Total	Percent (%) Yes	Percent (%) No
None of the above	50	46	96	52.1	47.9
LCSH	14	82	96	14.6	85.4
MeSH	10	86	96	10.4	89.6
TGN	9	87	96	9.4	90.6
ITIS	8	88	96	8.3	91.7
NBII	7	89	96	7.3	92.7
Not sure	7	89	96	7.3	92.7
EnvThes and/or LTER	6	90	96	6.3	93.7
GO	3	93	96	3.1	96.9
UAT	3	93	96	3.1	96.9
AGROVOC	1	95	96	1.0	99.0
ERIC	1	95	96	1.0	99.0
NALT	0	96	96	0	100

Table 12. Controlled Vocabularies Used by Data Curators and Repository Developers: Answers Supplied by Participants



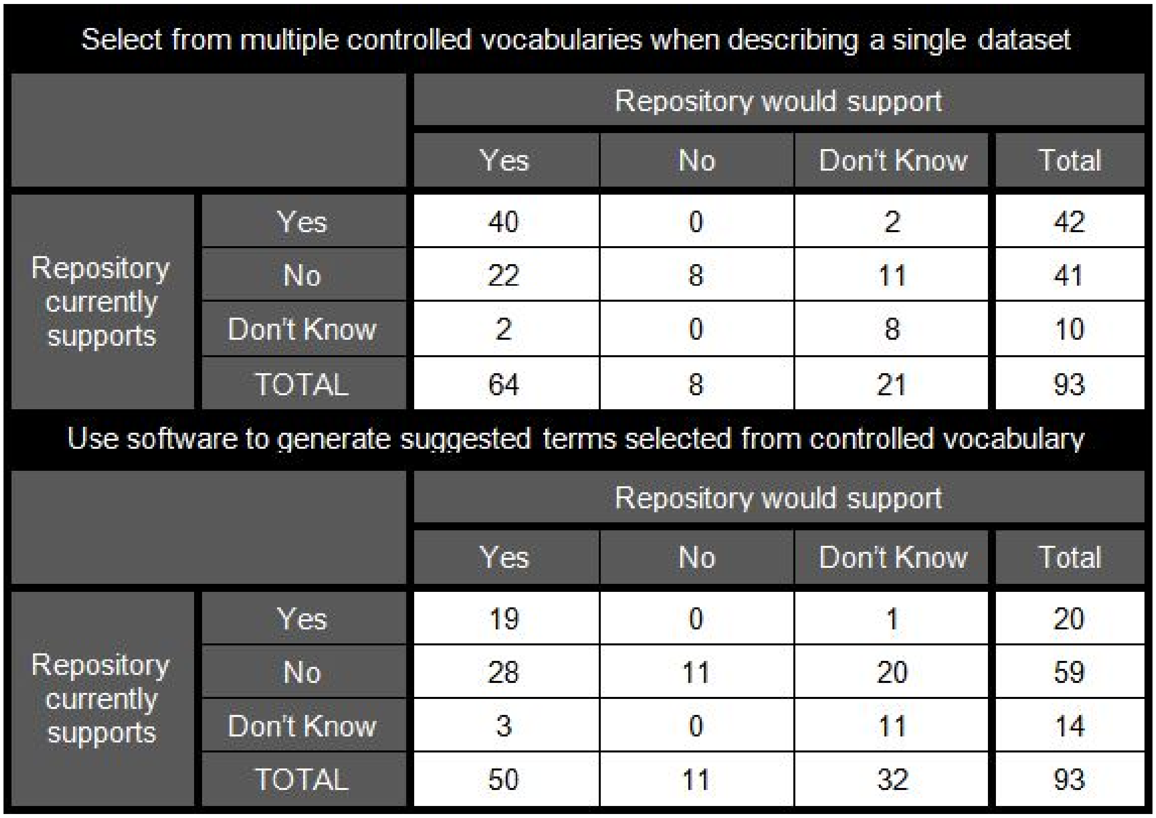
	Frequency	Percent (%)
NASA GCMD Earth Science Keywords	13	12.4
NetCDF CF (Climate and Forecast) Metadata Convention	11	10.5
ISO 19115 Topic Categories	7	6.7
CUAHSI Controlled Vocabularies	3	2.9
Dublin Core	3	2.9
GEMET (GEneral Multilingual Environmental Thesaurus)	3	2.9
Art and Architecture Thesaurus (AAT)	2	1.9
BioPortal Ontologies	2	1.9
DarwinCore	2	1.9
Ecological Metadata Language (EML)	2	1.9
FOAF	2	1.9
GeoNames	2	1.9
INSPIRE Spatial Data Themes	2	1.9
ISO 3166 Country Codes	2	1.9
SPASE (meta)Data Model	2	1.9
USGS Geographic Names Information System	2	1.9
BODC (British Oceanographic Data Centre) Parameter Discovery Vocabulary (P021)	1	1.0
CC (Colon Classification)	1	1.0
CEDB (Civil Engineering Database) Subject Headings	1	1.0
CoL (Catalogue of Life)	1	1.0
Custom Controlled Vocabulary	1	1.0
DataCite Kernel	1	1.0
DDC (Dewey Decimal Classification)	1	1.0
DDI (Document Data Initiative) Vocabularies	1	1.0
Dublin Kernel	1	1.0

	Frequency	Percent (%)
EU Publications Office Register of Corporate Bodies	1	1.0
EU Publications Office Register of Countries	1	1.0
EU Publications Office Register of Languages	1	1.0
EU Publications Office Register of Places	1	1.0
EUROVOC	1	1.0
FAST	1	1.0
FITS	1	1.0
Global Names Index (GNI)	1	1.0
IANA Media Types	1	1.0
ICPSR Thesaurus	1	1.0
INSCRIPTION Wordlists	1	1.0
Inspec Classification	1	1
InterNano Taxonomy	1	1
IVOA (International Virtual Observatory Alliance) Controlled Vocabularies	1	1.0
JACS (Joint Academic Coding System)	1	1.0
LC NACO Authority File	1	1.0
LC TGM (Thesaurus for Graphic Materials)	1	1.0
MMI (Marina Metadata Interoperability) Controlled Vocabularies	1	1.0
National Phenology Network Phenophase Definitions	1	1.0
NIST controlled vocabulary for security assessment, testing, and certification	1	1.0
OGC (Open Geospatial Consortium) geoSPARQL	1	1.0
Open Annotation Ontology	1	1.0
PO (Plant Ontology)	1	1.0
ProQuest Subject Categories	1	1.0
RCUK (Research Councils United Kingdom) Research Classifications	1	1.0



	Frequency	Percent (%)
SWEET (Semantic Web for Earth and Environmental Terminology) Ontology	1	1.0
TDWG Life Science Identifiers	1	1.0
THESAGRO	1	1.0
TRY Plant Trait Tables	1	1.0
US NODC (National Oceanographic Data Center) Controlled Vocabularies	1	1.0
USGS Geographic Names Information System	1	1.0
VSO (Virtual Solar Observatory)	1	1.0
W3C Provenance Ontology (PROV-O)	1	1.0
W3C Time Ontology	1	1.0
WMO (World Meteorological Organization) Sea Ice Nomenclature	1	1.0
TOTAL	105	100.0

Table 13. Crosstabulation of Controlled Vocabulary Actions Performed by Data Curators and Repository Developers



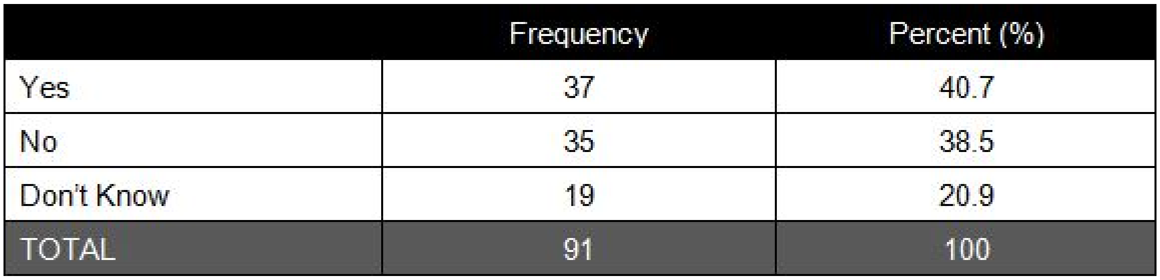
Select from multiple controlled vocabularies when describing a single dataset

		Repository would support			
		Yes	No	Don't Know	Total
Repository currently supports	Yes	40	0	2	42
	No	22	8	11	41
	Don't Know	2	0	8	10
	TOTAL	64	8	21	93

Use software to generate suggested terms selected from controlled vocabulary

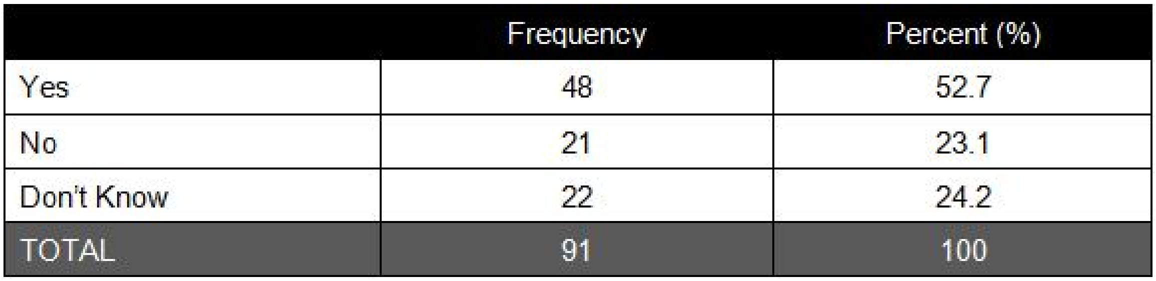
		Repository would support			
		Yes	No	Don't Know	Total
Repository currently supports	Yes	19	0	1	20
	No	28	11	20	59
	Don't Know	3	0	11	14
	TOTAL	50	11	32	93

Table 14. Use of Controlled Vocabularies Whose Terms Are Represented as URIs



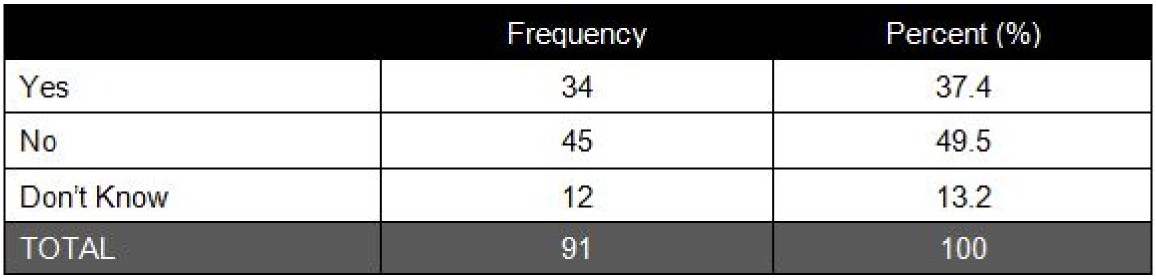
	Frequency	Percent (%)
Yes	37	40.7
No	35	38.5
Don't Know	19	20.9
TOTAL	91	100

Table 15. Use of Controlled Vocabularies Whose Terms are Not Represented as URIs



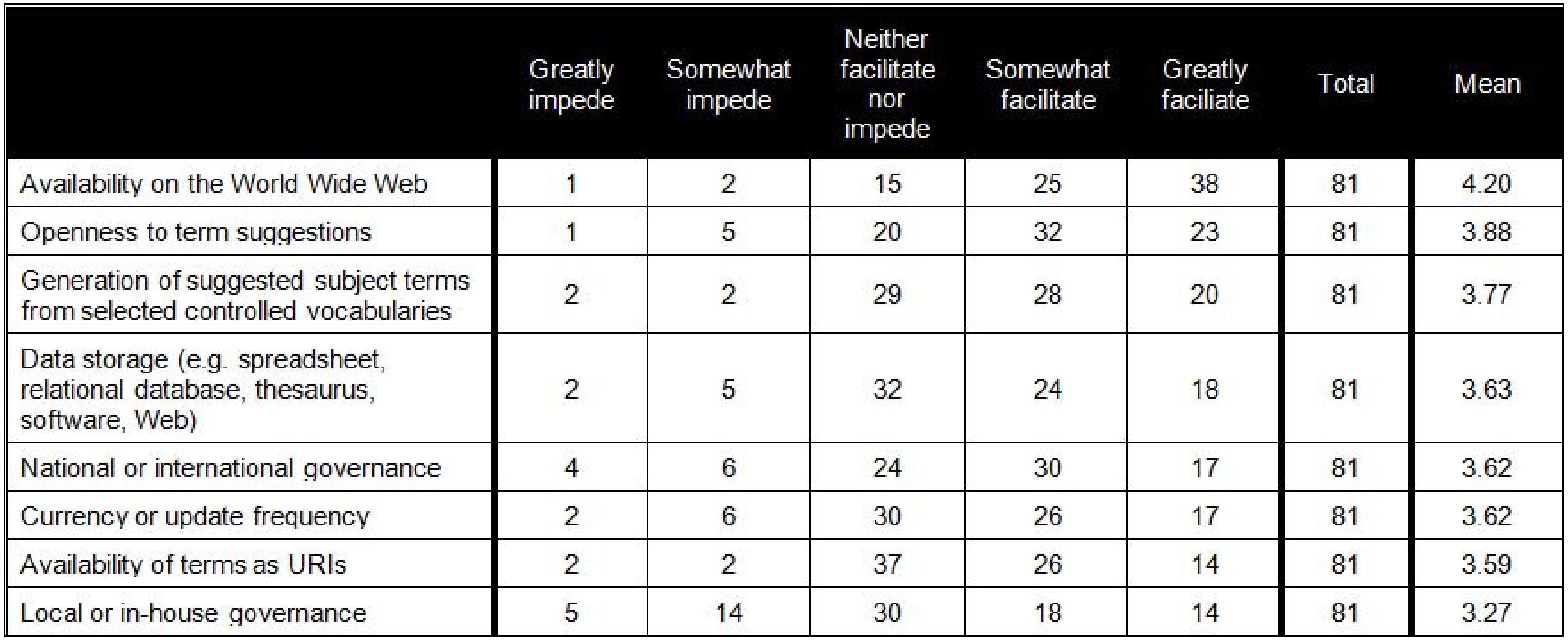
	Frequency	Percent (%)
Yes	48	52.7
No	21	23.1
Don't Know	22	24.2
TOTAL	91	100

Table 16. Validation of Terms against Specific Controlled Vocabularies



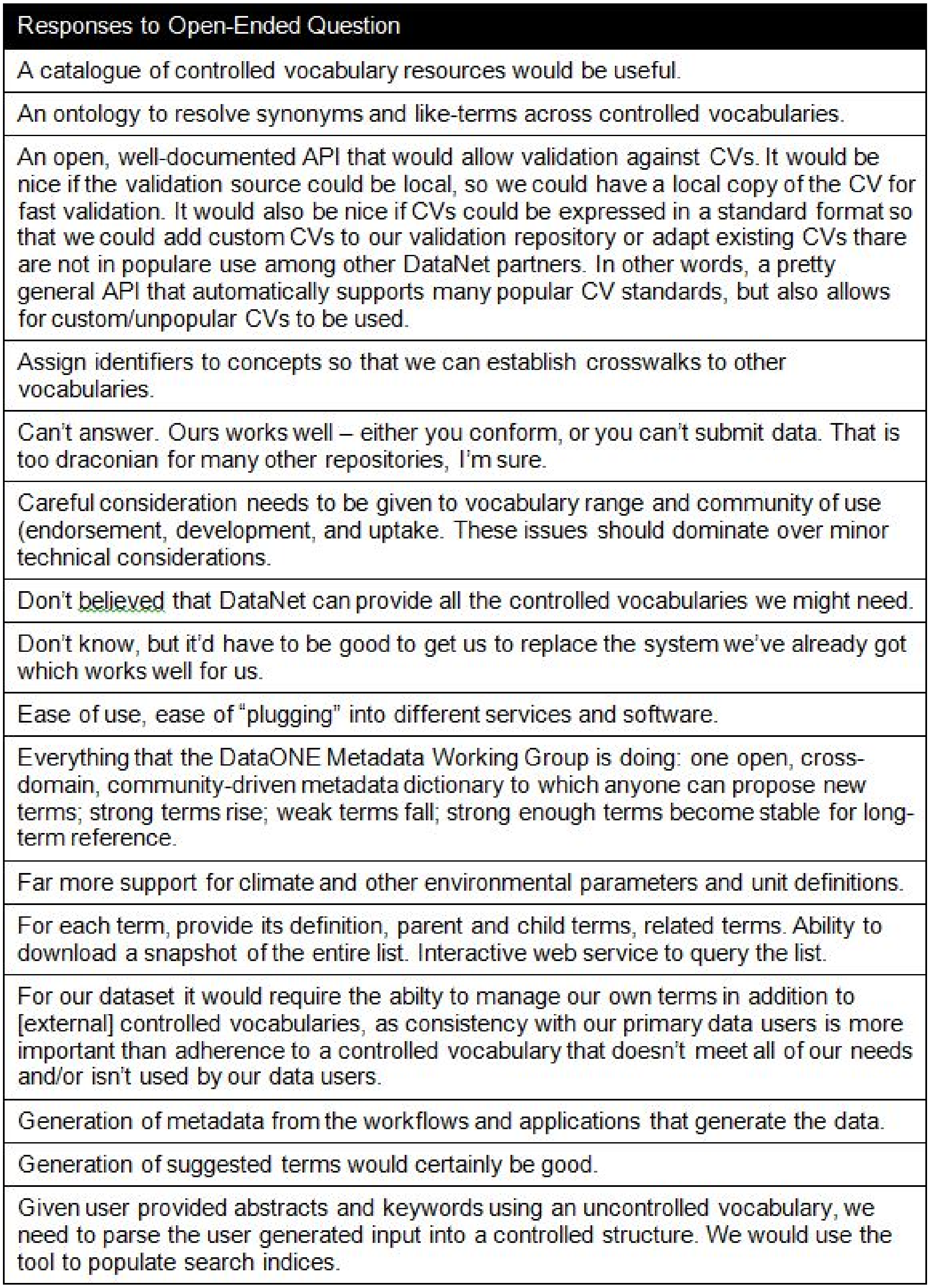
	Frequency	Percent (%)
Yes	34	37.4
No	45	49.5
Don't Know	12	13.2
TOTAL	91	100

Table 17. Facilitators and Inhibitors of Controlled Vocabulary Use



	Greatly impede	Somewhat impede	Neither facilitate nor impede	Somewhat facilitate	Greatly faciliate	Total	Mean
Availability on the World Wide Web	1	2	15	25	38	81	4.20
Openness to term suggestions	1	5	20	32	23	81	3.88
Generation of suggested subject terms from selected controlled vocabularies	2	2	29	28	20	81	3.77
Data storage (e.g. spreadsheet, relational database, thesaurus, software, Web)	2	5	32	24	18	81	3.63
National or international governance	4	6	24	30	17	81	3.62
Currency or update frequency	2	6	30	26	17	81	3.62
Availability of terms as URIs	2	2	37	26	14	81	3.59
Local or in-house governance	5	14	30	18	14	81	3.27

Table 18. Desired Tool Features



Responses to Open-Ended Question
A catalogue of controlled vocabulary resources would be useful.
An ontology to resolve synonyms and like-terms across controlled vocabularies.
An open, well-documented API that would allow validation against CVs. It would be nice if the validation source could be local, so we could have a local copy of the CV for fast validation. It would also be nice if CVs could be expressed in a standard format so that we could add custom CVs to our validation repository or adapt existing CVs thare are not in populare use among other DataNet partners. In other words, a pretty general API that automatically supports many popular CV standards, but also allows for custom/unpopular CVs to be used.
Assign identifiers to concepts so that we can establish crosswalks to other vocabularies.
Can't answer. Ours works well – either you conform, or you can't submit data. That is too draconian for many other repositories, I'm sure.
Careful consideration needs to be given to vocabulary range and community of use (endorsement, development, and uptake. These issues should dominate over minor technical considerations.
Don't believed that DataNet can provide all the controlled vocabularies we might need.
Don't know, but it'd have to be good to get us to replace the system we've already got which works well for us.
Ease of use, ease of "plugging" into different services and software.
Everything that the DataONE Metadata Working Group is doing: one open, cross-domain, community-driven metadata dictionary to which anyone can propose new terms; strong terms rise; weak terms fall; strong enough terms become stable for long-term reference.
Far more support for climate and other environmental parameters and unit definitions.
For each term, provide its definition, parent and child terms, related terms. Ability to download a snapshot of the entire list. Interactive web service to query the list.
For our dataset it would require the abilty to manage our own terms in addition to [external] controlled vocabularies, as consistency with our primary data users is more important than adherence to a controlled vocabulary that doesn't meet all of our needs and/or isn't used by our data users.
Generation of metadata from the workflows and applications that generate the data.
Generation of suggested terms would certainly be good.
Given user provided abstracts and keywords using an uncontrolled vocabulary, we need to parse the user generated input into a controlled structure. We would use the tool to populate search indices.

Responses to Open-Ended Question
I need mappings between controlled vocabularies for different communities.
I would use such a tool to add preferred terms to records while keeping free-text tags in place.
Ideally, disambiguation between similar terms with different usages and differing terms with similar semantics.
In my personal observations, science researchers are not as familiar with the jargon of "controlled vocabularies" and "ontologies." They need a tool that helps them connect the correct subject headings or keywords to their work, regardless of what scheme it is. They mostly don't care if it's LCSH or NBII – they just want the correct terms attached to their dataset. Do what you need to facilitate that.
Interoperable with many systems, many disciplines represented, deprecation tracking, backwards compatibility with older CV versions.
▪ It would be very useful to facilitate queries ▪ It would facilitate synthesis studies ▪ It would help archive migration and forward compatibility efforts ▪ It might enable cross-archive retrospective information science research
It would require availability of vocabulary's terms as Uniform Resource Identifiers (URIs).
Javascript UI tools to select from a DataNet controlled vocabulary.
▪ Mapping support ▪ Versioning and provenance support to both terms and mappings
Most important is to provide easy access to well defined and updated terms. Next, it is valuable to provide an open forum to propose new terms, improved definitions and a governance structure for approving mature new terms or modifications.
My "wish list" includes: ▪ Allow selection of specific vocabularies to be used in specific contexts ▪ Web services that support identification of candidate terms based on metadata content ▪ Tools for addressing shared terms in different vocabularies I would use the tool to facilitate keyword selection for datasets.
Propose new "candidate terms," commenting on "candidate terms," finalizing terms into a vocabulary, revisions. Replace informal email conversations.
Quickly training people who are unfamiliar with the vocabulary being used.
Registries, ontology mapping, annotation. Would use it to map between concepts and to describe limitations of those mappings.
Some level of ontology mapping between overlapping vocabularies is necessary. Ontologies need to be driven by practical use by data generators and data users within particular domains. A DataNet tool needs to be something that easily expands beyond DataNets and which facilitates the use of existing vocabularies, particularly at the data generation stage.

Responses to Open-Ended Question
Submittal of a parameter name to generate a set of controlled terms from vocabularies.
Such a tool would need to make it easy to find vocabularies of relevance, to identify terms within the vocabulary, and to identify cases where different terms are used across vocabularies.
Sync of DataNet terms with a simplified set of tables in a relational DB and a triple store.
The ability to be able to search or browse vocabularies from within a data deposit form or system.
The list is long. See http://marinemetadata.org/semanticframework for a start. Open source libraries and well defined procedures are likely to be as important as "a tool."
This question would require pages to answer. A big help would be a web service that semantically crosswalked across vocabularies.
This sounds like a very 1970s project in which term mapping across controlled vocabularies was attempted in the population or family planning area. Didn't work then, it is unlikely to work now because terms are already defined within their context, and it's the definitions, and not the terms, which do not cross boundaries.
To connect data between repositories.
Tool for inclusion of context sensitive suggestions when many values are applicable, tools that show the concrete value and benefit of using CV for that effort (this may be out of scope but important)…there is the stick of using CV but not enough carrots i.e. what would happen if you correctly used CV.
Tools are needed to assist scientists or curators with assigning terms to data. Tools should: ▪ Generate suggested terms ▪ Make suggested terms available via web services (for automatic insertion) ▪ Present forms to the user to reject some suggested terms (for manual insertion)
Tools that supported standardization.
User testing!
We might need discipline categories, especially if the scope of DataNet partners broadens to more than environmental, biological, ecological areas.
We would be more likely to use the tool if it was offered in the form of a web services API as opposed to a website or a desktop application. Web services would make the tool platform-independent and easier to embed within our current suite of software application.
Web services
Web-based capabilities to identify terms.

Appendix B: IRB Support

Title of Study

Advancing Interoperability of NSF DataNet Partners Through Controlled Vocabularies (IRB No. 13-1472)

Summary

The purpose of this study is to gather information about controlled vocabularies in use by the NSF DataNet Partners and other data repositories; purposes these controlled vocabularies serve; and both facilitators and inhibitors of controlled vocabulary use by different data repository stakeholders.

Participants will include data contributors, data curators, NSF DataNet Partner administrators, and repository infrastructure developers affiliated with NSF DataNet Partners and other data repositories.

The survey uses role associated with data repository to determine the question path. Some questions are directed at all participant communities e.g. knowledge of selected controlled vocabularies. In addition, a series of questions presented to those who describe data (either data contributors or data curators) differs from another series presented to those who make administrative decisions (data curators, NSF DataNet Partner administrators, and repository infrastructure developers).

Description of Risks

Risks are limited to breach of confidentiality. No sensitive subjects are included in the survey. The responses would present minimal to no risks to participants if divulged outside the research.

Consent Process

Participants will be required to provide electronic verification of a voluntary consent form before proceeding with the web survey.

Confidentiality of the Data

At the end of the survey, participants will have the option to provide name and email adress for possible follow-up. If participants choose to provide name and email address, these identifiers will be connected to the survey data indirectly through codes stored in a separate location from the survey data.

Appendix C: Survey Invitation & Reminder Templates

Survey Invitation Template

SUBJECT: Please participate (very brief survey!) for data contributors, curators, administrators, repository developers

The following survey examines controlled vocabulary use and challenges.

The survey is for data contributors, curators, administrators, and/or repository developers.

Completing the survey takes approximately 15 minutes (or less) to complete. To complete the survey, please click the following link: https://unc.qualtrics.com/SE/?SID=SV_3fU0xOeRbH6jntb.

NOTE: If you are unable to click on the link directly, please type the entire link into the address or location field at the top of your web browser, and press the ENTER key on your keyboard to access the survey.

The survey is supported by a supplement to the original NSF DataNet grant to DataONE in order to explore controlled vocabulary use within and across a broad spectrum of data repositories, including but not limited to the U.S. DataNet initiatives.

Sincerely,
Chelcie Rowell

--
Chelcie Rowell
Research Assistant, Metadata Research Center
School of Information and Library Science
University of North Carolina at Chapel Hill
chelcie@live.unc.edu | 770.862.0750

Survey Reminder Template

SUBJECT: REMINDER: Please participate (very brief survey!) for data contributors, curators, administrators, repository developers

Thanks to those who have already participated in this survey. We're eager for more participation.

Please participate if you have not yet completed this survey, and please feel free to forward this call to other lists and colleagues.

The following survey examines controlled vocabulary use and challenges.

The survey is for data contributors, curators, administrators, and/or repository developers.

Completing the survey takes approximately 15 minutes (or less) to complete. To complete the survey, please click the following link: https://unc.qualtrics.com/SE/?SID=SV_3fU0xOeRbH6jntb.

NOTE: If you are unable to click on the link directly, please type the entire link into the address or location field at the top of your web browser, and press the ENTER key on your keyboard to access the survey.

The survey is supported by a supplement to the original NSF DataNet grant to DataONE in order to explore controlled vocabulary use within and across a broad spectrum of data repositories, including but not limited to the U.S. DataNet initiatives.

Sincerely,
Chelcie Rowell

--
Chelcie Rowell
Research Assistant, Metadata Research Center
School of Information and Library Science
University of North Carolina at Chapel Hill
chelcie@live.unc.edu | 770.862.0750

Appendix D: Survey Instrument

Consent to Participate in a Research Study

Title of Study: Advancing Interoperability of NSF DataNet Partners Through Controlled Vocabularies (IRB No. 13-1472)

Principal Investigator: Chelcie Rowell | chelcie@live.unc.edu | 770.862.0750

Faculty Advisor: Jane Greenberg | janeg@email.unc.edu | 919.962.8366

What is the purpose of this study? To gather information about the use of controlled vocabularies to advance interoperability among National Science Foundation (NSF) DataNets.

Who is conducting this study? This study is being conducted by Chelcie Rowell, a Research Assistant with the Metadata Research Center at the School of Information and Library Science at the University of North Carolina at Chapel Hill.

Who should take part in this study? Individuals associated with any NSF DataNet Partner as well as scientists, curators, administrators, and repository developers involved in the deposit or management of scientific research data in repositories.

What will happen if I take part in this study? Participating in this study will take approximately 10–15 minutes of your time. You will be asked to complete a Web survey about your use of controlled vocabularies to describe scientific research data. Your decision to participate or decline participation in this study is completely voluntary and you have the right to terminate your participation at any time without penalty. If you do not wish to complete this survey simply close your browser.

What are the risks of participating in this study? There are no risks to individuals participating in this survey beyond those that exist in daily life.

How will my privacy be protected? At the end of the web survey, if you would be interested in being contacted for follow-up, you will have the option to provide contact information. If you choose to provide contact information, this identifying information will be stored separately from the survey data.

What if I have questions about this study? If you have questions about this research study, you may contact Chelcie Rowell by email at chelcie@live.unc.edu or by phone at 770.862.0750. If you have questions or concerns about your rights as a research subject you may contact, anonymously if you wish, the Office of Human Research Ethics at the University of North Carolina at Chapel Hill by phone at 919.966.3113 or by email at IRB_Subjects@unc.edu.

Q1 I have read and understand the above consent form, I certify that I am 18 years old or older and, by selecting the "I consent" answer choice, I indicate my willingness voluntarily to take part in this research study.

1 I consent

2 I do not consent

Determining Question Path

Q2 In the past twelve months, which of the following actions have you performed? Select all that apply.

1 Deposited research data with a data repository

2 Managed research data deposited with a data repository

3 Served as a PI, co-PI, or full-time employee of an NSF DataNet Partner

4 Developed systems, software, or other infrastructure to support a data repository

5 Other action related to a data repository [Please specify]
{SHORT TEXT RESPONSE}

Participants may select more than one answer choice for Q2. If answer choice 1 is selected for Q2, then the question block Questions for Data Contributors is shown. If answer choice 2 or 4 is selected for Q2, then the question block Questions for Data Curators and Developers is shown. If answer choice 2, 3, or 4 is selected for Q2, then the question block Questions for Data Curators, Administrators, and Developers is shown.

Questions for Data Contributors

Q3 A **controlled vocabulary** is a carefully selected list of terms that is used to describe resources (such as documents or datasets) so that they may be more easily retrieved by a search. Types of controlled vocabularies include **term lists**, **authority files**, **classification systems**, **thesauri**, and **ontologies**. The organization governing a controlled vocabulary makes decisions about what terms are included as well as decisions about vocabulary storage, vocabulary editing, and vocabulary maintenance.

Q4 From which of the following controlled vocabularies have you selected subject terms when describing data deposited with any repository? Select all that apply.

1 AGROVOC (thesaurus of the Food and Agriculture Organization of the United Nations)

2 EnvThes (Environmental Thesaurus) and/or the United States LTER (Long Term Ecological Research Network) Vocabulary

3 ERIC (Education Resources Information Center) Thesaurus

4 GO (Gene Ontology)

5 ITIS (Integrated Taxonomic Information System)

6 LCSH (Library of Congress Subject Headings)

7 MeSH (Medical Subject Headings)

8 NALT (National Agricultural Library Thesaurus)

9 NBII (National Biological Information Infrastructure) Biocomplexity Thesaurus

10 TGN (Thesaurus of Geographic Names)

11 UAT (Unified Astronomy Thesaurus)

12 None of the above

13 Not sure

Q5 Please list any additional controlled vocabularies from which you have selected subject terms when describing data deposited with any repository.

{LONG TEXT RESPONSE}

Q6 How frequently have you selected subject terms from a controlled vocabulary in order to describe your research data deposited in any repository?

1 Never

2 At least once

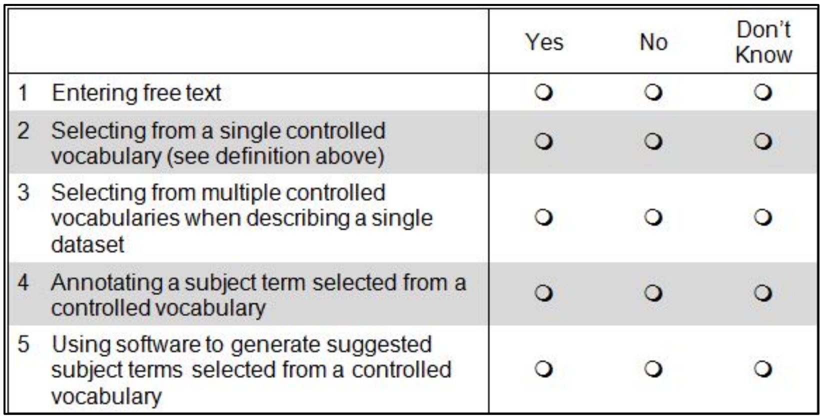
3 1–3 times per year

4 3–6 times per year

5 7+ times per year

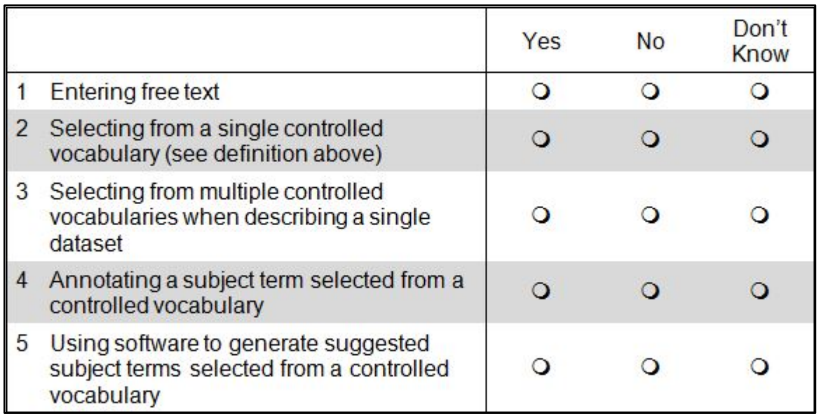
6 Other [Please specify] {SHORT TEXT RESPONSE}

Q7 Which of the following actions related to providing subject terms have you performed?



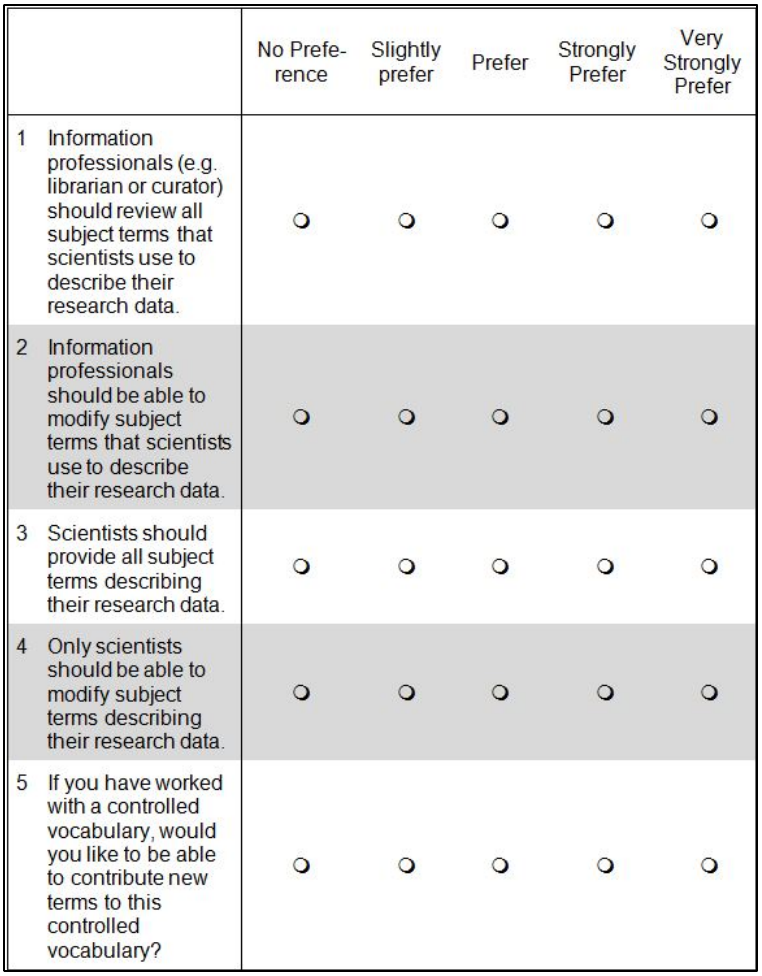
	Yes	No	Don't Know
1 Entering free text	○	○	○
2 Selecting from a single controlled vocabulary (see definition above)	○	○	○
3 Selecting from multiple controlled vocabularies when describing a single dataset	○	○	○
4 Annotating a subject term selected from a controlled vocabulary	○	○	○
5 Using software to generate suggested subject terms selected from a controlled vocabulary	○	○	○

Q8 If it were possible now, would you make use of the following functions in the next twelve months?



		Yes	No	Don't Know
1	Entering free text	○	○	○
2	Selecting from a single controlled vocabulary (see definition above)	○	○	○
3	Selecting from multiple controlled vocabularies when describing a single dataset	○	○	○
4	Annotating a subject term selected from a controlled vocabulary	○	○	○
5	Using software to generate suggested subject terms selected from a controlled vocabulary	○	○	○

Q9 Please indicate your preference for describing data deposited with any repository.



		No Prefe-rence	Slightly prefer	Prefer	Strongly Prefer	Very Strongly Prefer
1	Information professionals (e.g. librarian or curator) should review all subject terms that scientists use to describe their research data.	○	○	○	○	○
2	Information professionals should be able to modify subject terms that scientists use to describe their research data.	○	○	○	○	○
3	Scientists should provide all subject terms describing their research data.	○	○	○	○	○
4	Only scientists should be able to modify subject terms describing their research data.	○	○	○	○	○
5	If you have worked with a controlled vocabulary, would you like to be able to contribute new terms to this controlled vocabulary?	○	○	○	○	○

Q10 Do you believe it is important that scientists provide subject terms describing their own research data deposited in a repository?

1 Yes

2 No

If answer choice 1 is selected for Q10, then Q11 is shown.

Q11 Please indicate why it is important to you to provide subject terms describing your research data deposited in a repository. Select all that apply.

1 I know my discipline well.

2 I know how users are likely to search for my research data.

3 I like to control how my research data is represented.

4 Other [Please specify] {SHORT TEXT RESPONSE}

If answer choice 2 is selected for Q10, then Q12 is shown.

Q12 Please explain why it is not important to you to provide subject terms describing your research data deposited in a repository.

{LONG TEXT RESPONSE}

Questions for Data Curators and Repository Developers

Q13 A **controlled vocabulary** is a carefully selected list of terms that is used to describe resources (such as documents or datasets) so that they may be more easily retrieved by a search. Types of controlled vocabularies include **term lists**, **authority files**, **classification systems**, **thesauri**, and **ontologies**. The organization governing a controlled vocabulary makes decisions about what terms are included as well as decisions about vocabulary storage, vocabulary editing, and vocabulary maintenance.

Q14 Which of the following controlled vocabularies does your repository use? Select all that apply.

1 AGROVOC (thesaurus of the Food and Agriculture Organization of the United Nations)

2 EnvThes (Environmental Thesaurus) and/or the United States LTER (Long Term Ecological Research Network) Vocabulary

3 ERIC (Education Resources Information Center) Thesaurus

4 GO (Gene Ontology)

5 ITIS (Integrated Taxonomic Information System)

6 LCSH (Library of Congress Subject Headings)

7 MeSH (Medical Subject Headings)

8 NALT (National Agricultural Library Thesaurus)

9 NBII (National Biological Information Infrastructure) Biocomplexity Thesaurus

10 TGN (Thesaurus of Geographic Names)

11 UAT (Unified Astronomy Thesaurus)

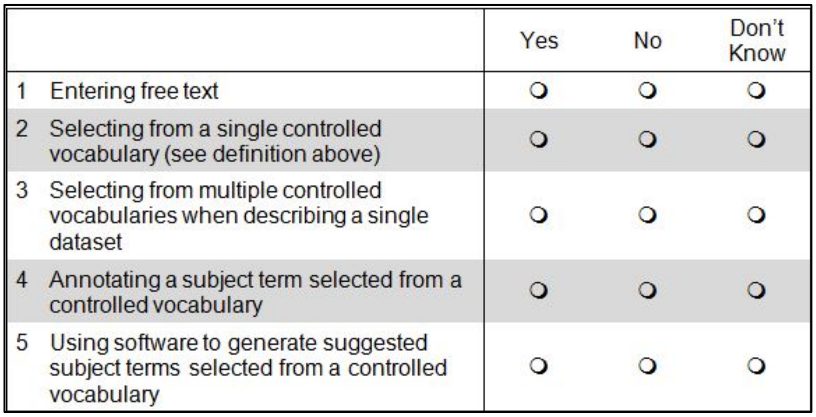
12 None of the above

13 Not sure

Q15 Please list any additional controlled vocabularies from which you have selected subject terms when describing data deposited with your repository.

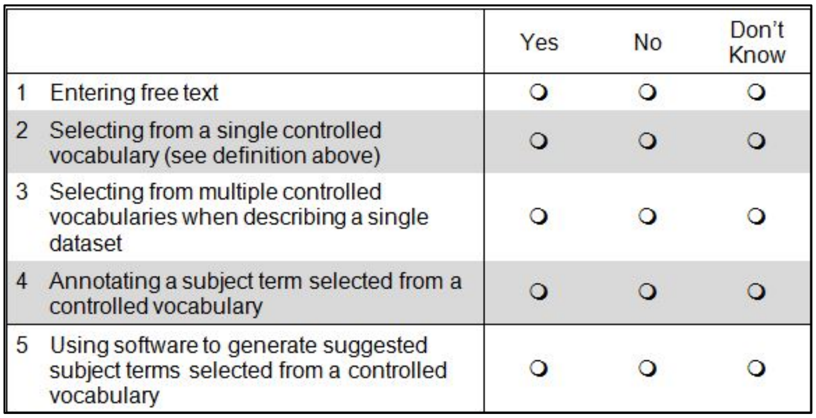
{LONG TEXT RESPONSE]

Q16 Which of the following functions related to providing subject terms does your repository support?



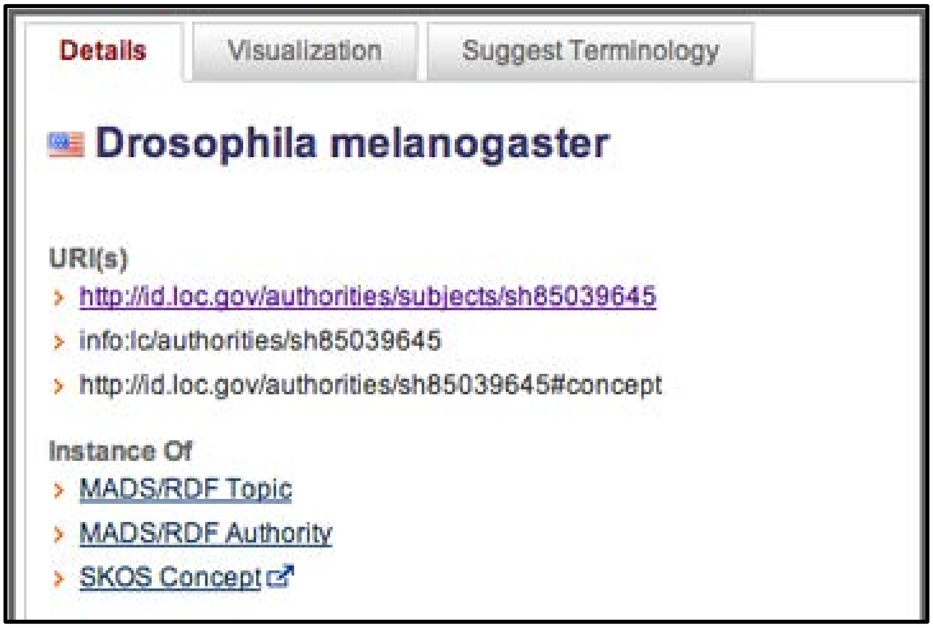
		Yes	No	Don't Know
1	Entering free text	○	○	○
2	Selecting from a single controlled vocabulary (see definition above)	○	○	○
3	Selecting from multiple controlled vocabularies when describing a single dataset	○	○	○
4	Annotating a subject term selected from a controlled vocabulary	○	○	○
5	Using software to generate suggested subject terms selected from a controlled vocabulary	○	○	○

Q17 If it were possible now, would your repository support the following functions in the next twelve months?



		Yes	No	Don't Know
1	Entering free text	○	○	○
2	Selecting from a single controlled vocabulary (see definition above)	○	○	○
3	Selecting from multiple controlled vocabularies when describing a single dataset	○	○	○
4	Annotating a subject term selected from a controlled vocabulary	○	○	○
5	Using software to generate suggested subject terms selected from a controlled vocabulary	○	○	○

Q18 Below is an example in which the term "Drosophila melanogaster" from Library of Congress Subject Headings (LCSH) is represented as the URI "http://id.loc.gov/authorities/subjects/sh85039645". Terms in a controlled vocabulary may or may not be represented as Uniform Resource Identifiers (URIs). A URI is a string of characters used to identify a name or a web resource.



Drosophila melanogaster

URI(s)

- http://id.loc.gov/authorities/subjects/sh85039645
- info:lc/authorities/sh85039645
- http://id.loc.gov/authorities/sh85039645#concept

Instance Of

- MADS/RDF Topic
- MADS/RDF Authority
- SKOS Concept

Q19 Does your repository make use of any controlled vocabularies whose terms are represented as Uniform Resource Identifiers (URIs)?

1 Yes

2 No

3 Don't Know

Q20 Does your repository make use of any controlled vocabularies whose terms are not represented as Uniform Resource Identifiers (URIs)?

1 Yes

2 No

3 Don't Know

Q21 Does your repository validate the subject terms selected by data contributors or data curators against any specific controlled vocabularies?

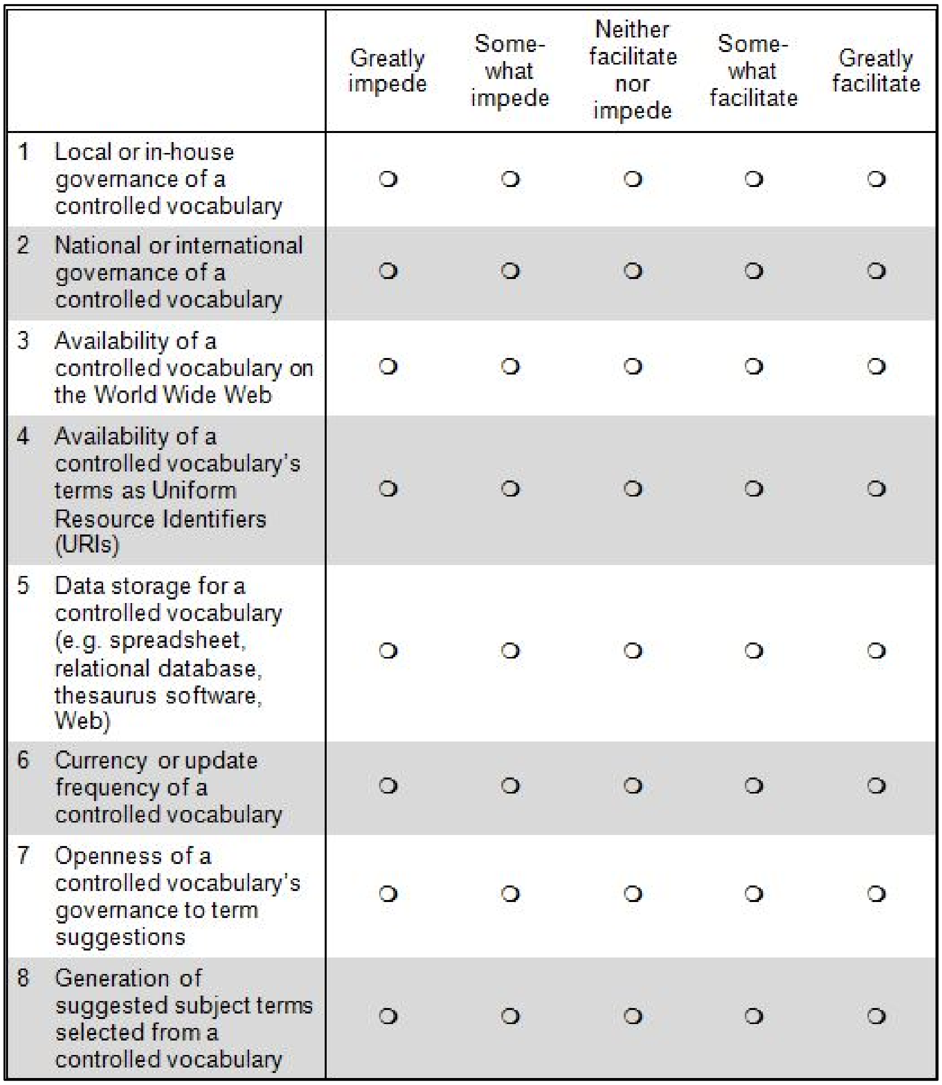
1 Yes

2 No

3 Don't Know

Questions for Data Curators, DataNet Administrators, and Repository Developers

Q22 Please rate how the following aspects or features facilitate or impede your use of controlled vocabularies to describe scientific research data.



	Greatly impede	Some-what impede	Neither facilitate nor impede	Some-what facilitate	Greatly facilitate
1 Local or in-house governance of a controlled vocabulary	○	○	○	○	○
2 National or international governance of a controlled vocabulary	○	○	○	○	○
3 Availability of a controlled vocabulary on the World Wide Web	○	○	○	○	○
4 Availability of a controlled vocabulary's terms as Uniform Resource Identifiers (URIs)	○	○	○	○	○
5 Data storage for a controlled vocabulary (e.g. spreadsheet, relational database, thesaurus software, Web)	○	○	○	○	○
6 Currency or update frequency of a controlled vocabulary	○	○	○	○	○
7 Openness of a controlled vocabulary's governance to term suggestions	○	○	○	○	○
8 Generation of suggested subject terms selected from a controlled vocabulary	○	○	○	○	○

Q23 Has participation in a DataNet or other data repository influenced your plans for using controlled vocabularies? How?

{LONG TEXT RESPONSE]

Q24 If a tool were to be built that would support the use of controlled vocabularies within and across DataNet Partners, what features would it need? How would you use such a tool?

{LONG TEXT RESPONSE]

Demographic Questions

Q25 Several NSF DataNet Partners are addressing infrastructure challenges for scientific research data. With which NSF DataNet Partner are you involved? Select all that apply.

1 DataONE (Data Observation Network for Earth)

2 DFC (DataNet Federation Consortium)

3 SEAD (Sustainable Environment Actionable Data)

4 TerraPop (Terra Populus)

5 Not applicable

If answer choice 1 is selected for Q25, then Q26 is shown.

Q26 How long have you been involved with DataONE (Data Observation Network for Earth)?

1 Fewer than 6 months

2 6 months to 2 years

3 2 years to 5 years

4 More than 5 years

If answer choice 2 is selected for Q25, then Q27 is shown.

Q27 How long have you been involved with the DFC (DataNet Federation Consortium)?

1 Fewer than 6 months

2 6 months to 2 years

3 2 years to 5 years

4 More than 5 years

If answer choice 3 is selected for Q25, then Q28 is shown.

Q28 How long have you been involved with SEAD (Sustainable Environment Actionable Data)?

1 Fewer than 6 months

2 6 months to 2 years

3 2 years to 5 years

4 More than 5 years

If answer choice 4 is selected for Q25, then Q29 is shown.

Q29 How long have you been involved with TerraPop (Terra Populus)?

1 Fewer than 6 months

2 6 months to 2 years

3 2 years to 5 years

4 More than 5 years

If answer choice 2, 3, or 4 is selected for Q2, then Q30 is shown.

Q30 With which DataONE member node(s) are you affiliated? Select all that apply.

1 Cornell Lab of Ornithology Avian Knowledge Network (AKN)

2 Dryad

3 Earth Data Analysis Center (EDAC)

4 Ecological Society of America (ESA) Data Registry

5 Knowledge Network for Biocomplexity (KNB)

6 Long Term Ecological Research Network (LTER)

7 Oak Ridge National Laboratory Distributed Active Archive Center

8 ONEShare

9 Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)

10 South Africa National Parks (SanParks)

11 United States Geological Survey (USGS) Core Sciences Clearinghouse

12 University of California Curation Center (UC3) Merritt Repository

13 Not applicable

If answer choice 2, 3, or 4 is selected for Q2, then Q31 is shown.

Q31 With which DFC (DataNet Federation Consortium) project partner(s) are you affiliated? Select all that apply.

1 Cyber-Infrastructure-Based Engineering Repositories for Undergraduates (CIBER-U) at Drexel University
(Engineering Data Grid)

2 The iPlant Collaborative
(Plant Biology Data Grid)

3 National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center
(Hydrology Data Grid)

4 Ocean Observatories Initiative
(Oceanography Data Grid)

5 The Odum Institute for Research in Social Science
(Social Science Data Grid)

6 Renaissance Computing Institute (RENCI)
(Hydrology Data Grid)

7 Temporal Dynamics of Learning Center
(Cognitive Science Data Grid)

8 University of South Carolina College of Engineering and Computing
(Hydrology Data Grid)

9 University of North Carolina at Chapel Hill Institute for the Environment
(Hydrology Data Grid)

10 Not applicable

If answer choice 1 is selected for Q2, then Q32 is shown.

Q32 Which of the following best describes your primary field of study? Select one.

1 Biology

2 Climatology

3 Cognitive Science

4 Computer Science

5 Ecology

6 Engineering

7 Environmental Science

8 Geography

9 Geoscience

10 Hydrology

11 Information and Library Science

12 Marine Science

13 Physics and Astronomy

14 Social Science

15 Mathematics

16 Other [Please specify]

If answer choice 1 is selected for Q2, then Q33 is shown.

Q33 Which of the following best describes your secondary field(s) of study? Select all that apply.

1 Biology

2 Climatology

3 Cognitive Science

4 Computer Science

5 Ecology

6 Engineering

7 Environmental Science

8 Geography

9 Geoscience

10 Hydrology

11 Information and Library Science

12 Marine Science

13 Physics and Astronomy

14 Social Science

15 Mathematics

16 Other [Please specify]

If answer choice 1 is selected for Q2, then Q34 is shown.

Q34 With which DataONE member node(s) have you deposited data? Select all that apply.

1 Cornell Lab of Ornithology Avian Knowledge Network (AKN)

2 Dryad

3 Earth Data Analysis Center (EDAC)

4 Ecological Society of America (ESA) Data Registry

5 Knowledge Network for Biocomplexity (KNB)

6 Long Term Ecological Research Network (LTER)

7 Oak Ridge National Laboratory Distributed Active Archive Center

8 ONEShare

9 Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO)

10 South Africa National Parks (SanParks)

11 United States Geological Survey (USGS) Core Sciences Clearinghouse

12 University of California Curation Center (UC3) Merritt Repository

13 Not applicable

If answer choice 1 is selected for Q2, then Q35 is shown.

Q35 With which DataNet Federation Consortium (DFC) data grid(s) have you deposited data? Select all that apply.

1 Cognitive Science Data Grid

2 Engineering Data Grid

3 Hydrology Data Grid

4 Oceanography Data Grid

5 Plant Biology Data Grid

6 Social Science Data Grid

7 Not applicable

Concluding Questions

Q36 Please share any additional comments about this survey or the topic of advancing interoperability of NSF DataNet Partners through controlled vocabularies.

{LONG TEXT RESPONSE}

Q37 If you would be interested in being contacted for follow-up, please provide your contact information below. If you choose to provide contact information, this identifying information will be stored separately from the survey data.

1 First Name {SHORT TEXT RESPONSE}

2 Last Name {SHORT TEXT RESPONSE}

3 Email Address {SHORT TEXT RESPONSE}

End of Survey Message (Consent)

Thank you for participating in this research study. Your participation and the participation of others will help us to understand existing use of controlled vocabularies across NSF DataNets and other data repositories as well as opportunities for interoperability among data repositories.

End of Survey Message (Non-Consent)

Thank you for considering participation in this research study. If you have questions about this research study, you may contact Chelcie Rowell by email at chelcie@live.unc.edu or by phone at 770.862.0750. If have questions or concerns about your rights as a research subject you may contact, anonymously if you wish, the Office of Human Research Ethics at the University of North Carolina at Chapel Hill by phone at 919.966.3113 or by email IRB_Subjects@unc.edu.