

5-2015

Toward a Commons of Geographic Data

Joseph J. Campbell
University of Maine

Follow this and additional works at: <http://digitalcommons.library.umaine.edu/etd>



Part of the [Geographic Information Sciences Commons](#), and the [Spatial Science Commons](#)

Recommended Citation

Campbell, Joseph J., "Toward a Commons of Geographic Data" (2015). *Electronic Theses and Dissertations*. 2261.
<http://digitalcommons.library.umaine.edu/etd/2261>

This Open-Access Dissertation is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DigitalCommons@UMaine.

TOWARD A COMMONS OF GEOGRAPHIC DATA

By

James J. Campbell

B.A. LeMoyne College, 1968

M.A. New York University, 1970

MLIS University of Wisconsin-Madison, 2006

A DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

(in Spatial Information Science and Engineering)

The Graduate School

The University of Maine

May 2015

Advisory Committee:

Harlan Onsrud, Professor of Spatial Informatics, Advisor

M. Kate Beard Tisdale, Professor of Spatial Informatics

Max Egenhofer, Professor of Spatial Informatics

Elisabeth Allan, Professor of Higher Education

Torsten Hahmann, Assistant Professor of Spatial Informatics

On behalf of the Graduate Committee for James J. Campbell, I affirm that this manuscript is the final and accepted dissertation. Signatures of all committee members are on file with the Graduate School at the University of Maine, 42 Stodder Hall, Orono, Maine.



April 30, 2015

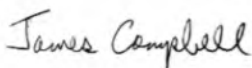
Dr. Harlan J. Onsrud, Professor of Spatial Informatics

Date

This a copyrighted work protected under Creative Commons Version 4 license:
CC-BY-NC. Terms of the license may be found
at <http://creativecommons.org/licenses/by-nc/4.0>.

LIBRARY RIGHTS STATEMENT

In presenting this dissertation in partial fulfillment of the requirements for an advanced degree at the University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for "fair use" copying of this dissertation for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this dissertation for financial gain shall not be allowed without my written permission.

Signature: 

Date: April 30, 2015

TOWARD A COMMONS OF GEOGRAPHIC DATA

By James Campbell

Dissertation Advisor: Dr. Harlan J. Onsrud

An Abstract of the Dissertation Presented
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
(in Spatial Information Science and Engineering)
May 2015

Making scientific data openly accessible and available for re-use is desirable to encourage validation of research results, and/or economic development. A significant body of spatially-referenced, locally-produced data produced by individual researchers, non-profit groups, private associations, small companies, universities, and non-governmental organizations across the United States is not online and therefore not generally available to professional scientists and to the general public. If there were an online environment, a “Commons of Geographic Data,” where that data could be deposited or registered, and where users could access and re-use it, what infrastructure characteristics might potential contributors find desirable in order for them to be willing to contribute their data without monetary compensation; and what infrastructure characteristics might potential users find desirable in order for them to be willing to access, investigate, and use such contributed data?

Based on data preservation literature, this study hypothesized three such potential characteristics as desirable. Using a combination of qualitative and quantitative methods, this study examined the desirability of these infrastructure capabilities in a non-statistical sample of potential contributors and potential users. The results of both the qualitative and quantitative research support the hypothesis. The results can provide guidance for those who may wish to design such a commons environment for locally-generated, spatially-referenced data in the future, and may also be of use to those that operate repositories of other types of data.

Acknowledgments

My very special thanks go out to my advisor, Prof. Harlan Onsrud, for his encouragement and support throughout my studies. I also appreciate the work of the members of my committee, and I am thankful for the good spirits, collegiality, and humane scholarship of the faculty who I have had the pleasure of knowing during my studies in the Spatial Informatics Program. They are, in my estimation, a model for how learning should be shared and knowledge moved forward.

TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	iv
LIST OF TABLES.....	xii
LIST OF FIGURES.....	xiii
Chapter	
1. INTRODUCTION.....	1
1.1. General Context.....	1
1.2. Study Questions.....	2
1.3. Chapter Contents.....	3
1.4. Motivation for This Study.....	3
1.4.1. The Concept of an Information Commons.....	4
1.4.2. A Commons of Geographic Information.....	4
1.4.3. Initiatives to Make Geographic Information Freely Available.....	5
1.5. Scope of This Study.....	6
1.5.1. A Commons of Geographic Data.....	6
1.5.2. Motivations to Participate in a Commons of Geographic Data.....	7
1.6. Research Hypotheses.....	9
1.7. Research Questions.....	10
1.8. Research Products.....	11

1.9. Research Approach.....	11
1.9.1. Research Process.....	12
1.9.2. Sources for Recruiting Potential Contributors and Users for This Study.....	14
1.9.2.1. Characteristics of Interviewees.....	15
1.9.3. Post Interview Processing.....	15
1.10. Dissertation Organization.....	16
2. SCIENTIFIC, SOCIAL, AND LEGAL CONTEXTS.....	18
2.1. The Enclosure of the Information Commons.....	18
2.2. Changes in Law and Technology.....	21
2.2.1. The Claim of Copyright.....	21
2.2.2. The Term of Copyright Protection.....	25
2.2.3. The Role of Government in Protecting Copyright.....	26
2.2.4. Scope of Copyright Protection.....	29
2.2.5. Summary of Changing Laws and Technology.....	33
2.3. Reactions to the Enclosure of the Information Commons.....	34
2.3.1. Legislate.....	34
2.3.2. Litigate.....	42
2.3.3. Legally Re-Interpret.....	46
2.3.4. Create Alternatives.....	53
2.3.4.1. GNU General Public License.....	54
2.3.4.2. The Street Performer Protocol.....	54
2.3.4.3. Creative Commons.....	56

2.3.4.4. Open Access Publishing.....	59
2.3.4.5. Open Access to Data.....	62
3. ACCESS TO SCIENTIFIC DATA IN THE 21ST CENTURY: RATIONALE AND ILLUSTRATIVE USAGE RIGHTS REVIEW.....	64
3.1. Introduction.....	64
3.2. The Role of Data in 21st Century Science.....	65
3.3. Reasons for Calls for Open Access to Scientific Data.....	66
3.3.1. Traditional Functions: Experiment Replication and Validation.....	67
3.3.2. Avoidance of Duplication.....	70
3.3.3. Access to Data as a Human Right.....	71
3.3.4. Data Preservation and Archiving.....	73
3.4. Desirable Characteristics of Data Collection and Storage Systems.....	75
3.4.1. Access.....	76
3.4.2. Clear Use Conditions.....	76
3.4.3. Findability.....	77
3.4.4. Evaluation Capability.....	78
3.4.5. Technical Characteristics.....	78
3.5. A Brief Overview Of Recent Initiatives To Provide Open Access To Scientific Data.....	80
3.5.1. Open Access Data Repository Growth.....	80
3.5.2. Access To U.S. Government Generated Data.....	82

3.5.3. Access To Data In The U.S. Generated By Non-Federal Government Bodies.....	83
3.5.4. Private and Corporate Initiatives.....	84
3.5.5. Non-U.S. Access Efforts.....	86
3.6. Usage Rights And Data Repositories: A Brief Review.....	87
3.7. Chapter Conclusion.....	94
4. POTENTIAL CONTRIBUTOR PERSPECTIVES ON DESIRABLE CHARACTERISTICS OF AN ONLINE DATA ENVIRONMENT FOR SPATIALLY-REFERENCED DATA.....	96
4.1. Introduction.....	96
4.2. Potential Contributor Motivation.....	98
4.3. Desirable Characteristics of Data Repositories.....	99
4.4. Hypothesis.....	100
4.5. Method.....	101
4.5.1. Interviewees and Data Types.....	101
4.5.2. Qualitative Data Collection Process.....	103
4.5.3. Quantitative Data Collection Process.....	104
4.6. Results and Discussion.....	106
4.6.1. Hypothesis Sub-part (a).....	107
4.6.1.1. Qualitative Findings.....	107
4.6.1.2. Quantitative Results.....	110

4.6.2. Hypothesis Sub-part (b).....	116
4.6.2.1. Qualitative Findings.....	116
4.6.2.2. Quantitative Results.....	118
4.6.3. Hypothesis Sub-part (c):.....	120
4.6.3.1. Qualitative Findings.....	120
4.6.3.2. Quantitative Results.....	121
4.6.4. Repository Maintenance.....	122
4.6.4.1. Qualitative Findings.....	123
4.6.4.2. Quantitative Results.....	124
4.7. Chapter Conclusions.....	125
4.7.1. Limitations.....	125
5. DESIRABLE CHARACTERISTICS OF AN ONLINE DATA COMMONS FOR SPATIALLY-REFERENCED, LOCALLY-GENERATED DATA FROM DISPARATE CONTRIBUTORS.....	128
5.1. Background.....	128
5.2. Volunteered Geographic Information.....	131
5.3. Desirable Characteristics of an Online Spatially Referenced Data Repository.....	132
5.4. Hypothesis.....	134

5.5. Method.....	135
5.5.1. Methodological Limitations.....	135
5.5.2. Interviewees and Data Types.....	136
5.5.3. Qualitative Data-collection Process.....	137
5.5.4. Quantitative Data-collection Process.....	138
5.6. Results.....	139
5.6.1. Hypothesis Subpart (a): Simple Clear Terms of Use.....	139
5.6.1.1. Qualitative Findings.....	139
5.6.1.2. Quantitative Results.....	140
5.6.2. Hypothesis Subpart (b): Search Mechanism.....	143
5.6.2.1. Qualitative Findings.....	143
5.6.2.2. Quantitative Results.....	144
5.6.3. Hypothesis Subpart (c): Peer Evaluation.....	145
5.6.3.1. Qualitative Findings.....	145
5.6.3.2. Quantitative Results.....	146
5.7. Chapter Summary and Conclusions.....	150
5.7.1. Limitations.....	150
5.7.2. Directions for Future Research.....	151
5.7.3. Possible Wider Applications.....	151

6. CONCLUSIONS AND RECOMMENDATIONS.....	153
6.1. Study Motivation.....	153
6.2. Study Goals.....	154
6.3. Conclusions Based on This Study.....	155
6.4. Recommendations for Information Architects of a Commons of Geographic Data.....	155
REFERENCES.....	158
APPENDIX: Data repository sites with usage rights referenced to the list above.....	167
BIOGRAPHY OF THE AUTHOR.....	184

LIST OF TABLES

Table 1.	Data repository sites with usage rights referenced.....	91
	to the list above	
Table 2.	Data repository sites with more complete description of usage rights.....	167

LIST OF FIGURES

Figure 1.	Attribution.....	111
Figure 2	Importance of Non-Commercial Only Use.....	112
Figure 3.	User Modification of Data.....	113
Figure 4.	Types of Data That Might Be Withheld.....	114
Figure 5.	Importance of Ability to Attach Descriptions.....	119
Figure 6.	Ability of Users to Comment on Suitability for Use.....	122
Figure 7.	Importance of Long Term Repository Maintenance.....	124
Figure 8.	Importance of knowing conditions for use for data.....	141
Figure 9.	Would any conditions prevent you from examining data?.....	142
Figure 10.	Importance of being able to search for data in different ways.....	144
Figure 11.	Importance of being able to comment on suitability of data for use.....	147
Figure 12.	Would comments of others affect your decision to examine data?.....	148
Figure 13.	Importance of using a screen name when commenting on data....	149

CHAPTER 1

INTRODUCTION

1.1. General Context

A significant body of spatially referenced locally produced data exists on the hard drives and back-up systems of individual researchers, schools, non-profit groups, private associations, small companies, and other non-governmental organizations across the United States. Examples include a faculty member or graduate student doing research in a non-Geographic Information Science (GIS) field such as geology or public health, research which requires the generation of a sizable amount of spatially referenced data; a high school class project that locates and catalogs all of the trees over ten feet tall in a small town; a homeowners' association that monitors the water quality and plant growth of the lake on which their property is located; or a local commercial medical supply service that has mapped all of the handicapped accessible entrances to buildings in its delivery area as part of its business process, rather than for any particular geographic research reason.

In all of these cases, the data gathered by these small local originators could be of great value to others—if its existence and provenance were known, if the data were available through repositories using standards-based metadata and search mechanisms, if the quality of the data were evaluated, and if the rights to its use were clear. At present, however, very little of this data is available to scientific researchers and other potential users: it is, for all intents and purposes, “invisible” or, at best, “partially visible.” While there are many efforts at the

national and state levels to make government generated spatially referenced data available to the public (see Chapter 3), no such effort exists to collect and make available this type of privately generated local data. In recent years, mapping products such as Google Earth have allowed individuals to add information to location points on maps but these services fall far short of desirable functionality in terms of clear legal rights information, standards-based metadata, provenance, and suitability for purpose peer evaluation.

One proposed solution to make this currently invisible or partially visible data available would be the creation of a Commons of Geographic Data (National Research Council (U.S.) Committee on Licensing Geographic Data and Services 2004, Onsrud et al. 2004) which would enable local private data generators who wished to do so to make their information readily findable and available in a publicly accessible environment so that others could make use of it.

1.2. Study Questions

A number of questions naturally arise when contemplating the design of a Commons of Geographic Data (CGD). This study will focus on two of them as a step toward defining specifications that can be used as part of an infrastructure design for a Commons of Geographic Data. Specifically, this study investigates:

- key factors that would help motivate private generators of generally non-publicly available spatially referenced data sets to be willing to contribute their data to a commons environment, and
- factors that would help motivate users of generally non-publicly available spatially referenced data sets to be willing to access and use that data.

1.3. Chapter Contents

This Introduction briefly outlines:

- the motivation for this study
- the scientific, policy, and legal contexts which gave rise to the proposal for the development of a Commons of Geographic Data
- the scope of the study, the research questions addressed, and hypotheses advanced in the context of a possible CGD
- the approach and methodologies used in this study, and
- the structure of this dissertation as a whole.

1.4. Motivation for This Study

Any attempt to actually obtain support for building a Commons of Geographic Data will need to deal with two important questions: (1) would anyone be interested in contributing to or using the data in such a repository; and (2) if so, what functional characteristics would such a repository need to have in order to help motivate contributors and users to actually make use of it.

The importance of the first question was driven home when a research proposal was submitted to a funding agency and was declined. One of the reasons that several reviewers cited was the uncertainty about whether a CGD would actually be used if it were created. One reviewer, paraphrasing the tag line from a popular film, *Field of Dreams*, asked “what makes you think that ‘If you build it, they will come’?”

This study is an attempt to help answer that question. Hopefully, having a research based response to that question will be useful for potential future efforts

to build such a commons. Just as importantly, identifying some of the infrastructure characteristics that would be attractive to potential data contributors and data users would be useful for the designers of any commons-type online environment that could serve as a repository for locally generated spatially referenced data, and might also be helpful to operators of existing repositories who wished to understand their users better.

1.4.1. The Concept of an Information Commons

Although there are many visions of what an information commons could be, they all share the principle that information placed in a commons environment should be available at no cost to anyone who wishes to use it without obtaining prior permission from the copyright owner, as long as any conditions of use the owner attaches to the information are respected by the user. Any material in the public domain qualifies as part of an information commons as does any material that the copyright owner chooses to make available for use under conditions which do not require prior permission from the owner, provided any stipulated conditions of use are adhered to. Creative Commons licenses are an example of “some rights reserved” conditions put on usage of copyrighted material but which allow use without obtaining prior permission.

1.4.2. A Commons of Geographic Information

The impetus to make scientific information available to researchers and to the general public extends to the community of geographic information sciences. For example, a Study Committee of the National Research Council recommended

that “The geographic data community should consider a National Commons in Geographic Information where individuals can post and acquire commons-licensed geographic data. The proposed facility would make it easier for geographic data creators (including local to federal agencies) to document, license, and deliver their datasets to a common shared pool, and also would help the broader community to find, acquire, and use such data. Participation would be voluntary.” (National Research Council (U.S.) Committee on Licensing Geographic Data and Services 2004) The meaning of “commons-licensed geographic data” in this case comports well with the general description of an information commons above.

Such a proposed Commons of Geographic Information would operate alongside today's commercial marketplace for geographic data and information.

1.4.3. Initiatives to Make Geographic Information Freely Available

Both in the U.S. and in other countries around the world, initiatives are underway to make large-scale geographic information freely available to scientists and to the general public in the spirit, if not in the form, suggested by the NRC. Many of these are reviewed in Chapter 3.

At present, however, there is no significant effort underway to make broadly available spatially referenced data which is generated by local sources for their own purposes, and which is not generally accessible to the public. Even if such data were exposed to some extent through services such as Google Maps, the lack of standards-based metadata, clarity of licenses and use options, and questions about provenance might limit its usefulness even if it were to be

discoverable. It is this body of data that a Commons of Geographic Data seeks to make widely available.

1.5. Scope of This Study

This study is focused on factors and functions that could assist in the eventual creation of a Commons of Geographic Data.

1.5.1. A Commons of Geographic Data

While a Commons of Geographic Information might deal with any type of spatially related information ranging from in-progress working papers to teaching materials to published peer-reviewed articles to finished maps to raw data sets, this study focuses on one type of spatially related information: data sets themselves. In particular, we focus on the type of locally-generated, currently “invisible” or “partially visible” data described above. The goal of this study is to understand the characteristics of an infrastructure in which the generators of such data might be willing to choose to place their data in a commons environment where it would be available to others who might wish to use it for their own purposes.

We therefore concentrate on elements that would contribute to the creation of an infrastructure for what we refer to as a Commons of Geographic Data. While such a CGD could, and hopefully will, become a component of a larger Commons of Geographic Information, for the purposes of this study a CGD is viewed as a stand alone infrastructure that, if constructed, could be used productively on its own (Onsrud et al. 2004).

In order to create an effective Commons of Geographic Data, there must be both willing contributors of data that would be freely available for use without seeking prior permission of the owner, and willing users of the contributed data. This simple fact leads to the focus of this study: what infrastructure functions or characteristics would motivate potential contributors and potential users to be willing to participate in a Commons of Geographic Data?

1.5.2. Motivations to Participate in a Commons of Geographic Data

Although existing literature does not directly address factors that might motivate potential contributors to make their data available in a commons context, we can reasonably postulate certain motivating factors based on limited evidence from GIS-related literature; from research on motivations for contributing to web sites such as Wikipedia or Flickr, and to myriad sites utilizing Google Earth and other contemporary geolocation tools. We may also draw on substantial evidence from the literature on open source software development, recommender systems, volunteerism in general, and general infrastructure requirements for the efficient operation of archival and other data storage and access systems.

For example, we know from the literature on open source software development that the amount of time necessary to complete software development tasks strongly influences the likelihood that a volunteer will undertake that task. Other things being equal, the greater the time requirement, the smaller the number volunteers who are willing to undertake the task (Hars and Ou 2002). It is reasonable to postulate that the CGD infrastructure should

put as few demands on a contributor's time as possible to maximize the number of contributors.

We also know from the open source movement, as well as from research into motivations for open access publishing (and academic publishing in general), that reputation and receiving credit for one's work are important factors for many who create (Lakhani and Wolf 2005, Weber 2004, Goodchild 2007). It is therefore reasonable to hypothesize that attribution for their contributions would be an issue of concern to potential CGD contributors, as would some level of control over how their contributions are used by others.

We know from research into user searching behavior that users are more likely to be satisfied when they can quickly find what they are looking for based on searching terminology that makes sense to them. We can reasonably postulate that this functionality would be of concern to potential users of the commons (Hearst et al. 2002).

We know from Internet phenomena such as the Amazon.com, Flickr, delicious, and Slashdot web sites, as well as the power law phenomenon exhibited by blog use (Shirky 2003), that people rely heavily on the opinions of peers for a wide variety of decision making purposes ranging from whether to buy a book, look at a photo, or take the time to read another's posting on a web site or access a blog (Barabasi 2002, Shirky 2003). We can reasonably postulate that a data review system for contributed data would increase use of that data, and might possibly be a requirement for use of the commons for many potential users.

1.6. Research Hypotheses

This study hypothesizes that:

1) the following components are important to motivate local spatially-referenced data owners to be willing to contribute data to a commons environment:

(a) a simple, clear licensing mechanism that includes, if the contributor chooses, an assurance that the owner would receive credit for the contribution, and would have the option to choose which usage rights the owner is willing to pass on to users and which usage rights the owner wishes to retain

(b) a simple process for attaching descriptions to the data. The contributor could choose “plain English” user descriptions rather than controlled vocabulary items. These would be processed by the system into standards-based metadata without requiring knowledge of metadata systems or controlled vocabulary terms on the part of the contributor; or the contributor could use controlled vocabulary terms if the contributor so chose

(c) a simple post-publication peer evaluation mechanism that will both provide feedback for contributors, and provide information on quality and suitability for use for users.

2) the following components are important to motivate potential users to be willing to use contributed data in a commons environment:

(a) a simple, clear licensing mechanism that reveals ownership of, and conditions for use of, the contributed data

(b) a simple, effective searching/ finding mechanism which provides an option to search using controlled vocabulary, “plain English” keywords, or both

(c) a simple post-publication peer evaluation mechanism that will provide feedback for contributors, and provide information on quality and suitability for purpose for users.

In the context of this study, the term “simple” carries three meanings. The first is the common-sense, everyday sense of not complicated. The second meaning indicates that a contributor or user of the types of data that might be placed in, or consulted in, a Commons of Geographic Data would need no special geo-disciplinary expertise in order to contribute or use data. The third sense refers to simple as requiring a modest time commitment. All three of those senses are included in the use of the term in this study.

The characteristics listed in the hypothesis above parallel those described as essential for data management and preservation cited in studies by scholarly organizations such as the *Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content* (Caplan et al 2003) and *To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering* (ARL 2006).

1.7. Research Questions

Given this set of hypotheses, we are faced with the following questions:
would this specific set of infrastructure functions be sufficient to:

(a) motivate potential local owners of spatially referenced data to be willing to contribute their data sets to a Commons of Geographic Data? (Some owners will never be interested in contributing data without monetary

compensation. These owners are therefore not considered potential contributors and are not included in this study.)

(b) motivate potential end-users to be willing to use these data sets, and to contribute to evaluating the data sets contributed? (Potential end-users are those who would consider using data in a commons environment. Some possible end-users may have philosophical or other reasons which would prevent them from using data in a commons environment. These end users are not included in this study.)

1.8. Research Products

Based on the outcome of this research with potential contributors and potential users, we identify a set of functions and characteristics that are motivationally important and should be considered for inclusion in the design of a system infrastructure for a Commons of Geographic Data. These functions might also be of use to others who design or operate data access systems.

1.9. Research Approach

The research process employed a standard combination of qualitative and quantitative methods through a series of sequential steps to test each component of the hypotheses and arrive at the study's conclusions. This overall research process integrates an analytic inductive approach with qualitative methodology (Creswell 2007). The interview instrument, generated codes, and other tools used in the qualitative portion of the study, as well as the questionnaire developed for

the quantitative portion of the study are available electronically in a supplementary materials package available at http://digitalcommons.library.umaine.edu/sie_studentpub/.

1.9.1. Research Process

(1) For determining whether the hypothesized infrastructure components are sufficient to motivate potential contributors to contribute data to the commons, the process consisted of the following steps:

(a) one-on-one in-person semi-structured interviews with 10 potential data contributors. Potential data contributors were drawn from groups with spatially related data interests (see Sources for Recruiting Potential Contributors below). Interviews were 60-90 minutes in duration and took place at the potential contributor's home, office, or other place chosen by the contributor. Interviews were recorded using unobtrusive audio equipment. Interviewees were given brief pre and post interview questionnaires to identify whether the interview questions or process altered their understanding of a commons environment, and therefore may have influenced their responses to interview questions

(b) analyzed findings using accepted qualitative analysis protocols: transcribed interviews from audio recordings using a professional external transcribing service, coded and analyze transcripts using TAMS Analyzer software developed at Kent State University. See further discussion of process below

(c) based on information gathered in interviews, a short online questionnaire was constructed to confirm/ disconfirm hypothesis findings generated through

qualitative methods. Notice of the online questionnaire and invitation to take the survey was distributed through groups with members who are likely to be potential geodata contributors and/or users (See Sources for Recruiting Potential Contributors below). The goal was to collect at least 100 valid completed responses

(d) analyze quantitative results

(e) formulate conclusions and make recommendations for the design of a Commons of Geographic Data infrastructure based on study results.

(2) For determining whether hypothesized infrastructure components are sufficient to motivate potential users to use data from the commons, the process consisted of the following steps:

(a) one-on-one in-person semi-structured interviews with 10 potential data users, who happened to also be potential data contributors. Potential data users were drawn from groups with spatially related data interests. (See Sources for Recruiting Potential Contributors below) Interviews were 60-90 minutes, took place at the potential user's home, office, or other place chosen by the user, and were audio recorded using unobtrusive equipment

(b) findings were analyzed using accepted qualitative analysis protocols: interviews were transcribed from the audio recordings by an external transcription service. Transcripts were coded and analyzed transcripts using TAMS Analyzer software developed at Kent State University.

(c) based on information gathered in the interviews, a short online questionnaire was constructed to confirm/ disconfirm hypothesis results generated through qualitative methods. Notice of the questionnaire was

distributed through groups with members who are likely to be potential geodata users (See Sources for Recruiting Potential Contributors below). The goal was to collect at least 100 valid completed responses

(f) analyze quantitative results

(g) formulate conclusions and make recommendations for the design of a Commons of Geographic Data infrastructure based on results.

In order to minimize contamination of results through the in-person interview process itself, interviewees were given short pre and post interview questionnaires to determine whether questions during the interview had any effect in changing their opinions.

1.9.2. Sources for Recruiting Potential Contributors and Users for This Study

In order to test the hypotheses described above, willing subjects who are potential contributors to, or users of, the CGD were necessary for conducting interviews, generating questionnaire responses, and testing the sufficiency of the data review system.

For the in-person interviews, subjects were initially drawn from:

- individuals belonging to or representing organizations listed in the Maine Environmental Monitoring and Assessment Program Index. There are well over 100 organizations represented. All have an interest in, and most generate, spatially related data, and few have any public outlet for their data at present;
- individuals belonging to the Maine GIS Users Group. This group has over 100 members all of whom have an interest in spatially related data, and many of whom work with geodata on a regular basis.

For the online questionnaire, respondents are drawn from those who have not been participants in the in-person interview components of the study, and who are members of groups represented in the Maine Environmental Monitoring and Assessment Program Index, the Maine GIS Users Group, and/or who are subscribers to the Maine Geolibrary listserv. In addition, the online questionnaire and a request for responses is distributed online through the GSDI listserv (4500+ participants) and the URISA listserv (3000+ participants).

1.9.2.1. Characteristics of Interviewees

Once some initial interviewees were identified, a “snowball” approach to adding to the interviewee list was folded in: some of the initial interviewees suggested others who might be interested in participating.

The group of 10 interviewees were made up of 7 males and 3 females. All were native born U.S. residents. Seven were from Maine, one each from Massachusetts, Pennsylvania, and North Carolina. There was no attempt made to develop a multi-cultural pool of interviewees. It is possible therefore that interviews with a pool of interviewees from other cultures might result in different findings than those presented in the following chapters.

1.9.3. Post Interview Processing

Once the recorded interviews were transcribed by an external transcribing service, the transcripts were compared to the audio recordings and any necessary corrections to the written transcripts made.

Initial codes were deductively generated based on the topics of the interview. Additional inductive codes were added during the analysis process as indicated by the content of the interviews.

Once all of the interviews had been coded once and a complete set of codes developed, all of the interviews were gone through again with the completed set of deductive and inductive codes developed through the initial processing in hand. The point of this process was to reduce inadvertent bias on the part of the coder since the author was the only one coding the interviews, and to ensure that all relevant material had been coded consistently.

Findings were based upon the coded themes that emerged from the analysis.

1.10. Dissertation Organization

This dissertation is designed to explore a possible new dissertation format option for degrees in Spatial Information Science and Engineering. Chapters 1-3 contain background information on several areas of study that establishes the need for the research work reported in Chapters 4-5. Chapters 3, 4, and 5 are each comprised of self-contained, stand-alone published papers. As stand-alone papers, they necessarily contain some of the background materials which would seem duplicative if they were not meant to stand on their own in different journals.

The remainder of this study is organized as follows:

Chapter 2: A review and discussion of the legal, political, and scientific context for a Commons of Geographic Data;

Chapter 3: A review and discussion of access to scientific data under open access or “some rights reserved” initiatives with a specific focus on what the terms “open” and “free” mean in the context of a user's ability to access and reuse scientific data, especially spatially referenced data. This material was published in *CODATA Science Journal* (Campbell 2014) and appears with minor additions in Chapter 3;

Chapter 4: A study of the motivations of potential contributors willing to contribute data to a Commons of Geographic Data: specifically, establish and test a hypothesis about important infrastructure elements that would motivate potential local data contributors to place their data in a commons environment. This material was published in *First Monday* (Campbell 2015) and appears with minor additions in Chapter 4;

Chapter 5: A study the motivations of potential users of data in a Commons of Geographic Data: specifically, establish and test a hypothesis about important infrastructure elements that would motivate potential users to use data contributed to a commons environment. This material was published in *URISA Journal* (Campbell & Onsrud 2015) and appears with minor additions in Chapter 5;

Chapter 6: Apply the results from these studies to specify desirable infrastructure characteristics for a Commons of Geographic Data.

CHAPTER 2

SCIENTIFIC, SOCIAL, AND LEGAL CONTEXTS

2.1. The Enclosure of the Information Commons

In the United States, copyright is a socially granted right, and establishes a “bargain” between creators and the larger society. Creators get an “exclusive Right” to exploit the value of their work in economic terms, and society gets the benefit of having that work available for everyone to use and, after “limited Times,” to build on directly. Historically,

Intellectual property protection in the United States has always been about creating incentives to invent. Thomas Jefferson was of the view that ‘inventions cannot, in nature, be a subject of property’; for him, the question was whether the benefit of encouraging innovation was ‘worth to the public the embarrassment of an exclusive patent.’ On this long-standing view, free competition is the norm. Intellectual property rights are an exception to that norm, and they are granted only when - and only to the extent that - they are necessary to encourage invention. The result has historically been intellectual property rights that are limited in time, limited in scope, and granted only to authors and inventors who met certain minimum requirements. On this view, the proper goal of intellectual property law is to give as little

protection as possible consistent with encouraging innovation (Lemley 2004).

For about the first 180 years of U.S. history, copyright law worked quite well to pursue this Constitutional goal. One reason for this general success was the law; another was the technology needed to violate copyright in any significant way.

Copyright comes into play only when a copy is made, and until recently, the technological burden for making copies was large, e.g., to make a significant number of copies of printed material, one needed a printing press. The technological burden was even higher for making copies of a film or of early television programs. In this technological environment, copyright law was relatively simple and worked relatively well in balancing the Founders' goal: "To promote the progress of Science and the useful Arts" through granting those who claimed copyright an exclusive right for "limited Times" to exploit the commercial potential of their creative labors. In an atmosphere in which the technological burden is heavy, the legal burden may be light.

In the U.S., historically, not all creators asserted copyright on their works. From 1800-1976, only about 25% of works were copyrighted. Copyright owners had to affirmatively renew their copyrights to extend the length of protection, and only about 3% chose to do so. This may be because about 97% of copyrighted works exhaust their commercial potential within five years (Lessig 2004).

From 1976 on, however, there has been a sea change in copyright law, and a parallel change in the technological environment in which copyrighted materials

are distributed. This confluence of changes in law and technology has led to a concern that the balance implicit in the copyright “bargain” has shifted, and that law and technology have now placed the rights of copyright owners far above those of users of copyrighted material, and of society as whole.

What has changed since 1976, and why is there such concern that this change adversely affects the information commons? For the purposes of this discussion, the “information commons” consists of any information which a potential user of that information does not have to obtain explicit prior permission to use.

“Information,” in this sense, encompasses creative as well as informative works expressed in any tangible medium, including digital media. It also includes data *per se*, including spatially-referenced data.

Information commons materials include any work in the public domain. Works in the public domain are free for anyone to use in any way. In the U.S., facts *per se* cannot be copyrighted, and so are in the public domain. However, arrangements of facts may be copyrightable and thus excluded from the public domain (see Section 2.2.4. below).

In addition to these public domain works, there are works that are under copyright but for which the copyright owners have given prior permission for use, usually under specific conditions. Most often those conditions include ensuring that the work is attributed to the creator, and, to a lesser extent, that the work is used for non-commercial purposes. The conditions attached to Creative Commons (CC) licenses are examples of these “some rights restricted” conditions of use. (See discussion of Creative Commons below.)

With this description of the information commons in place, we return to the question: What has changed since 1976, and why is there such concern that this change will adversely affect the information commons? To answer that question, we must briefly highlight recent key changes in law, in technology, and in the convergence of the two that give rise to concerns that the information commons is, in James Boyle's terminology, being "enclosed" (Boyle 2003). We will then look at responses to this perceived enclosure, with a focus on recent initiatives designed to make information available with limited or no use restrictions.

2.2. Changes in Law and Technology

Three elements in U.S. copyright law have changed in recent decades: (1) the necessity for claiming copyright; (2) the term of copyright protection; and (3) the role of government in protecting copyright in a digital environment. A fourth element, the scope of what copyright covers, is also under discussion in the U.S. We examine each element in turn.

2.2.1. The Claim of Copyright

Prior to 1978, those who wished to obtain copyright protection for a work had the affirmative obligation to register that work with the Registrar of Copyright and assert ownership in order to benefit from the protections available through copyright law. That requirement changed in the Copyright Act of 1976, which went into effect January 1, 1978. Since then, copyright exists the moment any original work is fixed in a tangible medium.

This change from an “opt-in” system of asserting copyright to an “opt-out” system has had a tremendous impact on the public’s ability to access and re-use created materials. There is now a presumption in the law that anything created since 1976 is under copyright, and that therefore permission must be acquired for any protected use of that material. However, there is no longer an obligation to register copyrighted material in a central repository, nor even to identify the creator on any tangible copy of the work. This can make it extremely difficult to even find out who the copyright owner is, let alone track the owner down to gain permission for use.

In 1930, to take one example, over 10,000 books were published. In 2000, 176 of those titles were still in print. In 2013, according to the International Publishers Association (2014), 304,912 books were published in the U.S., and they have copyright protection for at least 70 more years. Yet, historically, almost 97% of published works exhaust their potential for economic return within five years. That may be why, historically when copyright extensions had to be affirmatively applied for, only 3% of copyrighted works applied for and had their copyright protection extended (Lessig, 2004). Most created works had so little economic potential after a short period of time that their creators let them pass into the public domain after only one term of copyright protection, usually 14 years. Others who wished to use those works were then free to do so with no restrictions, or in the case of the small percentage of copyrighted works whose copyright was extended, potential users could easily ascertain who owned the copyright, and for how long the period of protection ran.

Contrast this with the situation which a potential user of a work created since 1978 faces. It is unlikely that the percentage of works with economic value beyond five years will magically increase in a dramatic fashion. Instead, it is likely that those works will simply cease to be published once their economic value hits the point of diminishing returns, as has always been the case. A user who would like to build upon a certain work 50 years hence (or even 10 years hence) may have no way of knowing who the copyright owner is or how to contact that owner to ask permission. In such a scenario, it is unlikely that a potential user will risk using the copyrighted work without prior permission, and will simply decide not to use the work at all.

On the face of it, this may not seem like a cultural calamity: after all, the future user can simply create something entirely new. However, even a moment's reflection will point out the potential harm to the larger culture that this situation can cause. For example, suppose that the public domain had not been available to the Disney company, or that the origin of certain stories – many fixed in a tangible form – had not been possible to ascertain but were theoretically under copyright protection. Would society have had any of the tremendously successful re-creations of Pinocchio, Aladdin, or dozens of other public domain stories in the film format that today's adults grew up with, or would today's children have access to animated versions of *The Velveteen Rabbit* and a host of other previously copyrighted works in the Rabbit Ears series?

The change in copyright from an “opt-in” system, in which a creator has to affirmatively claim copyright for a work, to an “opt-out” system in which a creator has to affirmatively decline automatic copyright protection, if that is

possible to do under statute at all, changes the presumptions under which works may be used by future creators, and imposes a burden that makes it likely that future creators will feel constrained for generations to come from using material in works created today.

Since the term of copyright in the U.S. is so long, and since it is no longer necessary to claim or register a copyright, the U.S. now faces a serious “orphan works” problem. Orphan works are works that are presumably under copyright but for which no copyright owner can be found to ask for permission to use a work. Not surprisingly, creators are very hesitant to use or build on orphan works for fear that, even after an extensive though unsuccessful effort was made to find the owner, the copyright owner may subsequently appear and sue under the terms of the copyright act, and the financial penalties could be very severe, up to \$150,000 per unauthorized use if the use was willful.

The problem has become so acute that in 2005, Senators Hatch and Leahy requested that the Registrar of Copyrights study the problem, take testimony, and issue a report. The Registrar did so and the report she issued contained a number of recommendations as well as suggested language for legislation to amend the copyright law to deal with the problem of “orphan works” (Registrar of Copyrights 2006). While bills have been submitted in Congress to make the use of orphan works less risky for subsequent users who make good faith efforts to find a copyright owner but are unable to do so, none have made much progress and orphan works remain a serious limitation for those who wish to use or build on past works.

2.2.2. The Term of Copyright Protection

Congress first granted copyrighted works protection for a period of 14 years. For most of U.S. history, this term length, augmented by a possible extension of another 14 years if applied for, was the norm.

That period began to grow in the 20th century: the term of copyright was extended 11 times in 40 years culminating in the Sonny Bono Copyright Term Extension Act (CTEA) of 1997. The CTEA extended the term of copyright to an author's lifetime plus 70 years; or, for works in which copyright is held by a company, for 90 years from publication or 120 years from creation if the work was not published. For a rock musician or a young author or a student researcher who creates a work at age 20 and lives an average lifespan, the work would be under copyright protection for about 130 years under current U.S. law.

Not surprisingly, some have questioned whether a term of protection of 100+ years constitutes a grant of protection "for limited Times" or "promotes the Progress of Science and the useful Arts" as the Constitution directs. How does protection that extends 70 years beyond the death of the creator provide an incentive for that creator to produce more works that will eventually be available to society as a whole? How does such an extended duration of limited use promote progress? In the eyes of some, this term of protection does not enhance the delicate balance of the copyright "bargain" between creators and society. Instead, it introduces an entirely new vision of copyright, one which basically replaces the vision of copyright as a social compact with a vision of private property as the highest good when it comes to "Writings and Discoveries".

Congress, the courts and commentators increasingly treat intellectual property not as a limited exception to the principle of market competition, but as a good in and of itself. If some intellectual property is good because it encourages innovation, they reason, more is better. The thinking is that creators will not have sufficient incentive to invent unless they are legally entitled to capture the full social value of their inventions. On this view, absolute protection may not be achievable, but it is the goal of the system (Lemley 2004).

The CTEA extension of copyright protection, and its philosophical implications, have been challenged in court, and the courts have essentially deferred to Congress in deciding what the proper definition of “limited Times” may be (Eldred v. Ashcroft 2003). As it stands, therefore, the length of copyright protection in the U.S. is that contained in the Sonny Bono Copyright Term Extension Act.

2.2.3. The Role of Government in Protecting Copyright

Until recently, enforcement of copyright has been a civil matter in which a copyright owner who felt her rights had been violated would sue the alleged violator for damages and other, usually injunctive, relief.

But digital changes everything.

Digital technology makes it possible to make a perfect copy – or a thousand copies – of a digitally encoded work and to distribute those copies widely at a

cost approaching zero. Copyright owners, particularly in the music and film industries, have appealed to Congress to protect their intellectual property in this changed environment. Congress, as well as the executive branch, has responded to their requests.

In the past decade, the U.S. government has taken a much more active role in copyright enforcement, and in some cases has extended the legal definition of copyright violation to the criminal realm. The major piece of legislation in this effort has been the Digital Millennium Copyright Act (DMCA). The DMCA prohibits providing tools or even information that would enable circumventing any type of technological protection, usually referred to as DRM or “Digital Rights Management,” devised by copyright owners to limit access to their digital works. Not only does the act allow anyone harmed by violation of the act’s provisions to sue, it also makes willful violation for profit a felony.

Departing from the traditional role of the U.S. government in copyright matters, the U.S. Congress has funded a specific section within the Justice Department to pursue violations of copyright, and bills submitted in several recent sessions of Congress, several passed by at least one house but not, as of this writing, yet law, would authorize the Justice Department to sue alleged copyright violators in civil court on behalf of copyright owners, essentially making the Justice Department, funded at taxpayer expense, a legal firm for private copyright owners. And in 2008, Congress passed the Prioritizing Resources and Organization for Intellectual Property Act (PRO-IP) of 2008 which, among other things, created a “Copyright Czar” in the Executive Branch, and dramatically increased penalties for copyright infringement.

There are differing opinions on Congress's original intent regarding circumvention of DRM protections for non-infringing uses under the DMCA, e.g., what latitude to allow users in exercising fair use rights. However, courts thus far, especially in *Universal Studios v. Corley* in 2000, have opted for an interpretation that very narrowly defines exceptions and views the DMCA as protecting all types of DRM for virtually all purposes. (Samuelson, 2003)

Other government actions, while not specifically focused on copyright per se, have also had a significant impact on the overall health of the information commons. Governments allocate scarce public resources such as spectrum space in broadcasting. They also regulate competition through anti-trust and similar regulation. The past two decades in the U.S., and, in fact, throughout the world, has seen an unprecedented increase in the concentration of copyright ownership and in the ownership of channels of distribution for copyrighted works due to changes in government regulations and/or policies.

The results have been dramatic. As of 2003 in the U.S., 80% of music for retail sale was distributed by five companies. 70% of the major radio markets were controlled by four companies. In 1996, no single entity owned more than forty radio stations. After the changes in regulation introduced by the Communication Act revisions of 1996, Clear Channel Communications owned more than 1300 stations by 2003 after the FCC relaxed ownership rules, although since then the company has been divesting low performing stations. Of the 91 "major" television networks (including cable), 80% are owned by six companies. In 1992, 70% of prime time network programming was independently produced. Since the FCC rescinded rules separating content and transmission ownership, 75% of

prime time programming is owned by the networks (all numbers Lessig 2003). Recent FCC decisions have relaxed concentration of ownership rules even further, allowing, among other things, newspapers and broadcast outlets in some markets to be owned by the same companies for the first time in U.S. history.

This concentration of ownership of copyrighted materials and of the channels to distribute those materials has significant repercussions on the information commons. Access to a large portion of culturally important copyrighted material now lies in the hands of a relatively few owners. Those owners are in a powerful quasi-monopolistic position to control use of that material through technologically enforced licensing provisions, provisions which often are at odds with traditional user rights such as fair use and first sale.

This situation is becoming more and more prevalent as more publications are being distributed in digital form. This is particularly noticeable in the world of academic journals where a decade long trend of consolidation has led to a half dozen large corporations controlling access to scholarly journals. Predictably, the price increase for scholarly publications taken as a whole have significantly exceeded the rate of inflation for over a decade at a time when library budgets have been generally decreasing. The result is more limited access to journals and scholarly works both on campus and off.

2.2.4. Scope of Copyright Protection

Under U.S. law, copyright protection can only be extended to works which exhibit some degree of originality. Simple facts or even the obvious arrangement of facts cannot be protected under U.S. copyright law. A simple alphabetical

listing of place names, for example, or an alphabetical list of names and telephone numbers does not reach the threshold of originality needed for copyright protection. (*Feist Publications, Inc. v. Rural Tel. Service Co.*, 499 U.S. 340 (1991)) The bar for that originality is not high. In the words of Justice Sandra Day O'Connor, it requires but "a modicum of creativity." Even so, facts *per se* are in the public domain in the U.S.

Similarly, any public records generated by the federal government do not fall under copyright protection since the Copyright Act specifically excludes the federal government itself from claiming copyright in materials it produces. This includes everything from weather reports to court decisions to data on water purity to testimony before congressional committees. All material generated directly by federal government employees is in the public domain. Under the Freedom of Information Act, access to some federal government generated information may be limited by concerns for security or other political considerations but may not be limited because of copyright ownership by government.

Recently, there have been efforts that would have the effect of eroding components of the public domain in the U.S. The European Union now includes databases of facts as works that can gain protection, either through copyright or through a *sui generis* designation, and similar bills have been introduced in the U.S. Congress in recent years (e.g., HR 3261, HR 3872 in the 108th Congress, and others since). If bills of this type were to be enacted into law in the future, facts collected and arranged in even obvious ways would fall under copyright protection in the U.S.

This type of *sui generis* database protection scheme in the European Union has not, in the view of the Royal Society, been a good thing for science.

Advances of technology and commercial forces have led to new IP legislation and case law that unreasonably and unnecessarily restrict freedom to access and to use information. This restriction of the commons in the main IP areas of patents, copyright and database right has changed the balance of rights and hampers scientific endeavour. In the interests of society, that balance must be rectified (Royal Society 2003).

Another effort that could have an enclosing effect on the information commons is the set of initiatives undertaken between 2001 and early 2009 to “privatize” many functions of the federal government in the United States. For example, federal agencies such as NOAA generate or purchase outright a great deal of geographic data which anyone is free to use for any purpose, commercial or non-commercial, without seeking permission.

Some government officials, for example former Senator Rick Santorum (King 2012), feel that the government should not be generating any geographic data itself (except for some military or security purposes), or putting it to any use which private enterprises might provide. Federal agencies, they claim, should obtain the data they need from private sources. Some vendors and some government officials feel that information obtained from vendors should be licensed rather than purchased, and that the vendor should retain copyright in the materials generated.

If this were to become the standard practice of the federal government, this, too, would remove a great deal of information from the public domain, much of it paid for or subsidized in some way by taxpayer dollars. For example, a public high school class studying water quality in a local lake might no longer be able to download data from the Environmental Protection Agency's web site and re-use it without paying a fee to the private sector vendor that licensed that data to the EPA; or, at the least, not use the (formally free of use restrictions or cost) data before obtaining prior permission from the private sector vendor.

These and other initiatives that would broaden the scope of what can be protected under copyright, if enacted, would combine with the automatic grant of copyright (whether desired or not), the extension of copyright term, and the expanding involvement of government in copyright enforcement to further limit the preservation and development of the information commons in the digital age.

In the United States, there are two factors under copyright law that provide some utility to users of copyrighted materials in the face of the expansion of the scope of copyright: First Sale and Fair Use.

First Sale simply means that once a person purchases a lawful physical copy of a work, e.g., a book, an academic journal, a CD, or other copyrightable work, the copyright owner no longer exerts any control over that copy. The purchaser may loan the item, give it away, or even sell it since in none of those transactions is a copy of the work made. It is the First Sale doctrine that makes libraries and video rental stores possible in the U.S. As more and more sales of books and music become digital, however, licensing rather than selling is becoming more common even for personal purchases, thus minimizing the effect of the First Sale

doctrine since no sale technically took place, although some scholars are beginning to argue that many licenses actually should be considered as sales, for example, those that allow unlimited use by the licensee (Asay 2013). Even the Registrar of Copyright has raised the question of whether a marketplace in which everything digital is licensed is the most desirable one for America's economic future (Pallante 2013). These are, however, still questions and courts to date have held that licenses trump First Sale rights.

Fair Use enables use of copyrighted materials for certain purposes without the copyright owner's permission. The difficulty with Fair Use from the perspective of potential users is that Fair Use is a defense, not a right. A user may cite Fair Use as a defense if a copyright owner sues for violation of copyright. Although there is a four element test to help determine whether a particular use constitutes Fair Use, no one really knows until a judge's gavel falls. Lawrence Lessig once joked that "Your Fair Use right is your right to hire a lawyer." Nonetheless, Fair Use does provide some elasticity in an otherwise tightly bound U.S. copyright law.

2.2.5. Summary of Changing Laws and Technology

Since the term of copyright is now so long, since DRM cannot be legally circumvented under the DMCA, and since the copyright holder can impose license conditions which restrict or remove traditional user rights under copyright law, such as fair use and first sale, and then enforce those license provisions through the use of DRM, Pamela Samuelson has suggested that DRM

might more accurately be described as “digital restrictions management” (Samuelson 2003).

And, indeed, that is the way that many view the current situation in copyright in the U.S.: as a situation in which law and technology have combined to radically alter the traditional balance between copyright owners and users of copyrighted materials in favor of copyright owners.

2.3. Reactions to the Enclosure of the Information Commons

As it became clear that the Sonny Bono Copyright Extension Act and the Digital Millennium Copyright Act were altering the copyright landscape in an unprecedented way in today’s digital environment, those who found this landscape alteration undesirable or unacceptable began to respond. Responses took a variety of forms and approaches to addressing the problem of “enclosure.” We classify the responses for the purposes of this review as:

- Legislate
- Litigate
- Legally re-interpret
- Create alternatives

We examine them in turn.

2.3.1. Legislate

In the U.S., no bills have been introduced in the Congress over the past decade designed to specifically counteract the automatic grant of copyright, or to shorten its statutory duration. There are however, examples of bills introduced to

mitigate the effects of the CTEA in recent sessions of Congress, and to temper the effects of DRM technologies that copyright owners are increasingly using to control access to their digital products, technologies which are protected from circumvention under the DMCA.

While it is unlikely at this time that the periods of copyright protection codified into law through the Sonny Bono Copyright Term Extension Act will be reduced, given both World Intellectual Property Organization treaty obligations and the tenor of Congress, some legislative initiatives would have ensured that only those works whose owners actually wish to utilize copyright over the full term provided in the CTEA would receive the full term of copyright protection.

Rep. Zoë Lofgren, for example, twice introduced The Public Domain Enhancement Act (108th and 109th Congresses). It would have required copyright owners who wish to continue to enjoy copyright protection to affirmatively assert their copyright after 50 years by paying a small registration fee of one dollar. Absent that assertion, copyright would expire after 50 years. While the bill attracted some co-sponsors, it was referred to a sub-committee of the House Judiciary Committee and went nowhere.

Another of the “enclosing” laws, the DMCA, has had a number of consequences which were not intended, according to testimony that led to passage of the act. Some businesses, for example, have attempted to use threats of suits or prosecution based on Sections 1201(a)(1), 1201(a)(2), and 1201(b) of the DMCA to stifle reporting of shortcomings in their products (e.g., HP and Microsoft). Others, such as SONY, have attempted to stifle competition, and Lexmark invoked the DMCA in suing and actually obtaining an injunction

against Static Control Components, a company that sold aftermarket cartridges for Lexmark printers. (Lexmark International, Inc. v. Static Control Components, Inc., 387 F.3d 522 (6th Cir. 2004)) That injunction stood for almost a year before being vacated in October of 2004 by the Sixth Circuit Court of Appeals, which later also ruled against Lexmark’s DMCA violation claims. The process took years to conclude and had a large impact on innovation and competition in the printer ink industry until it was resolved.

This example, others like it, and examples of the DMCA being applied against consumers in ways that do nothing to thwart large scale digital “piracy,” which was Congress’s avowed intent in passing the DMCA, alarmed some in Congress, and led to the introduction of bills that were intended to rectify some of the imbalances that the sponsors felt the DMCA has created in favor of copyright owners.

Rep. Zoë Lofgren, for example, twice introduced the BALANCE Act, in 2003 and 2005, which is designed to make legal in the digital realm what has been – and remains – a user’s legal rights under copyright law in the paper realm. In proposing remedies, the bill’s summary at its first introduction in 2003 actually serves as a description of what its supporters believe has been lost to copyright users of digital materials under DRM protected by the DMCA:

Benefit Authors without Limiting Advancement or Net
Consumer Expectations (BALANCE) Act of 2003 - Amends
Federal copyright law to: (1) include analog or digital
transmissions of a copyrighted work within fair use
protections; (2) provide that it is not a copyright

infringement for a person who lawfully obtains or receives a transmission of a digital work to reproduce, store, adapt, or access it for archival purposes or to transfer it to a preferred digital media device in order to effect a non-public performance or display; (3) allow the owner of a particular copy of a digital work to sell or otherwise dispose of the work by means of a transmission to a single recipient, provided the owner does not retain his or her copy in a retrievable form and the work is sold or otherwise disposed of in its original format; and (4) permit circumvention of copyright encryption technology if it is necessary to enable a non-infringing use and the copyright owner fails to make publicly available the necessary means for circumvention without additional cost or burden to a person who has lawfully obtained a copy or phonorecord of a work, or lawfully received a transmission of it (HR1066, 108th Congress).

This proposed legislation, according to its sponsors, makes traditional fair use and first sale rights available in the digital domain, and would allow a user who has lawfully obtained a copy of a digital work to defeat DRM restrictions which interfered with exercising those rights. The bill, in plain language, got nowhere in either the 108th or 109th Congress. In its absence, courts have continued to rule that any kind of copy made in the process of transfer, even if only one copy exists

at the end of the process, is a violation of copyright (*Capitol Records, LLC v. ReDigi, Inc.* 2013).

In response to FCC efforts early in the first decade of the 21st century to mandate, at the behest of large content providers, that hardware manufacturers include a “broadcast flag” capable of preventing the copying of a video program that contained a “no copy” software code within it, former Senator Sam Brownback, now Governor of Kansas, introduced a very ambitious bill, the Consumers, Schools, and Libraries Digital Rights Management Awareness Act of 2003 (S. 1621). It had three goals: (1) to prevent the FCC from mandating that manufacturers build DRM detection technology into digital hardware such as computers, audio and video recorders, etc.; (2) to prohibit the sale of any such equipment without warning labels indicating how the technology could restrict consumer use of the product; and (3) to prohibit Internet Service Providers (ISPs) from being “compelled to make available to a manufacturer of a digital media product the identity or personal information of a subscriber or user of its service for use in enforcing the manufacturer's right relating to the use of such product” (S. 1621, 108th Congress). The bill drew no co-sponsors. It was read and sent to the Senate Committee on Commerce, Science, and Transportation, where it died.

Each of the provisions of the bill addressed what was seen at the time, September, 2003, as a serious potential or actual problem limiting the exercise of traditional consumer rights or traditional practices of law, and subsequent events have proved that assessment to be accurate: that is to say, all of what Senator Brownback identified as potential problems have occurred.

In the first instance, the FCC did, in fact, mandate that a “broadcast flag” be built into any hardware device capable of receiving a digital broadcast signal. These devices range from digital television sets to video cards in computers. The “broadcast flag” is an information bit which would signal to the hardware device that an instruction was coming which contained DRM restrictions, and that these should be implemented by the hardware device. Using broadcast flag DRM, for example, a content provider could allow a broadcast signal to be viewed but not recorded for later time shifted viewing, which would essentially make moot the *SONY v. Universal Studios* decision of the Supreme Court. (That 1983 decision allowed video recording of a broadcast signal in a private setting for personal use.) In short, the broadcast flag gave broadcasters wide latitude in the type of controls that they may unilaterally impose on users without even the fig leaf of a click-through contract.

A court challenge ensued and the “broadcast flag” was ruled invalid in 2005 by the Court of Appeals for the District of Columbia (*American Library Association et al v. Federal Communications Commission and United States of America*). That, however was not the end of the story. Several bills, e.g., the Communications, Consumer's Choice, and Broadband Deployment Act of 2006, attempted to give the FCC the authority to implement a “broadcast flag” although none of those bills have, to date, succeeded.

Nonetheless, the effect of this aborted effort has still been felt in the information economy. Microsoft’s Windows Media Center software, for example, was shown in 2008 to still be respecting “broadcast flag” code put into programs by content owners that prevented copying or time shifting allowed under the

SONY Betamax decision. Since this was not a government-imposed decision, it was perfectly legal behavior on Microsoft's part, although it still had the effect of controlling the use of information, in this case, television programming.

This example of Microsoft voluntarily implementing hardware/software control of what users can do with content was inspired, at least in part, by the numerous attempts in Congress to mandate DRM control mechanisms in hardware/software control of playback devices. While none of the bills has been fully successful, several attracted co-sponsors and garnered high profile hearings. Those bills would have imposed a requirement that all digital hardware, whether capable of receiving a broadcast signal or not, have DRM detection circuitry built in, circuitry which was acceptable to copyright owners, mainly the Recording Industry Association of America (RIAA) and the Motion Picture Association of America (MPAA). In the wake of these hearings, although the bills never passed, they did have the effect of forcing copyright owners and hardware manufacturers to enter into discussions of how to "voluntarily" implement DRM on the hardware level. And in 2010, the Federal Communications Commission granted cable and satellite television providers authority to use Selectable Output Control technology to essentially disable analog outputs on consumer set-top boxes when the content providers made movies still in theaters available via satellite or cable distribution. Brownback's bill, as it turned out, was speaking to a real issue in the limitation of access and use of information through DRM controls.

The third concern of Brownback's 2003 bill proved similarly prescient.

Under the DMCA, as interpreted by copyright owners, copyright owners could swear that someone was violating their copyright via the Internet, and then obtain an administrative order signed by a court clerk with no judicial review. That order would compel Internet Service Providers to provide the identity of customers who were identified by the RIAA only by an IP address. The third provision of the Consumers, Schools, and Libraries Digital Rights Management Awareness Act of 2003 addressed that topic and would have nullified that practice. The bill was not successful in that attempt. While some courts have since held that copyright owners must go through traditional judicial processes, including showing probable cause for action, in seeking to obtain a court order that would compel an ISP to turn over the identities of its customers, that is not a universal practice by any means.

Rep. Rick Boucher introduced a bill much more limited in scope than Brownback's bill, the Digital Media Consumers Rights Act (HR 107, 108th Congress), which would have addressed not hardware itself but products used in hardware, specifically Compact Discs. The bill would have simply mandated that copyright owners who produce compact discs for sale, and who include DRM controls on those discs that limit the way a purchaser can use the discs, must clearly label their products as containing such controls. While this approach does not directly remedy any limitations to traditional user rights under copyright law, it at least takes a "let the market decide" stance by providing consumers with information they need to make a market driven decision about which controls they are willing to live with. While having perfect market information is a crucial element of the economic theory underlying capitalism, and may seem to

be common sense in a free market economy, Rep. Boucher's fellow Congress people apparently did not find that to be the case. The bill never got out of committee.

In short, in the U.S., no legislative initiatives to ameliorate the effects of changes in law and in technology as they affect access to information have had any success up to the beginning of the first session of the 115th Congress while several, such as the PRO-IP Act (see 2.2.3 above) and the Fair Copyright in Research Works Act (see 2.3.4 below) move strongly in the opposite direction.

2.3.2. Litigate

While some were pursuing legislative remedies, others felt that recent changes in copyright law violated the spirit and letter of the U.S. Constitution. They mounted legal challenges to provisions of both the CTEA and to the Copyright Act of 1976, which made copyright protection automatic.

In *Eldred v. Ashcroft*, the lead plaintiff, Eric Eldred, made available on his web site, and in other fashions, works that had entered the public domain. Some of those works had their copyright terms extended retroactively by the CTEA. Eldred asserted he had standing in the case since his work and livelihood was directly impacted by the CTEA. He claimed in the suit that the CTEA was unconstitutional (1) because it violated the "limited Times" clause in the Constitution, and (2) because it constrained free speech.

The case went all the way to the Supreme Court, where it lost by a 7-2 vote. The majority found that the Constitution granted Congress the duty to determine what "limited Times" meant, and that the Court should defer to Congress's

judgment. Justice Breyer, one of the dissenters, had long argued against the extension of copyright: "Taken as a whole, the evidence now available suggests that, although we should hesitate to abolish copyright protection, we should equally hesitate to extend or strengthen it" (Breyer 1970), and he continued that argument in his dissent.

On the free speech issue, the Court held that the act did not change the "traditional contour of copyright," and that any free speech concerns raised by the act could be dealt with through copyright's traditional established safeguards, e.g., fair use.

While those who sought to have the CTEA declared unconstitutional failed to achieve that goal, others felt that elements of the Supreme Court's Eldred decision strengthened the case for asserting that a combination of recent changes in copyright law did, in aggregate, affect the "traditional contour of copyright" for a certain class of works, and therefore that these laws essentially created a situation which required "further first amendment scrutiny."

That is the approach taken by plaintiffs in *Kahle v. Ashcroft* (original name: case as decided is *Kahle v. Gonzales* 2007):

In this case, two archival organizations asked the U.S. District Court for the Northern District of California to hold that statutes that extended copyright terms unconditionally - the Copyright Renewal Act and the Copyright Term Extension Act (CTEA) - are unconstitutional under the Free Speech Clause of the First Amendment, and that the Copyright Renewal Act and CTEA together create an

“effectively perpetual” term with respect to works first published after January 1, 1964 and before January 1, 1978, in violation of the Constitution’s Limited Times and Promote...Progress Clauses. The Complaint asks the Court for a declaratory judgment that copyright restrictions on orphaned works - works whose copyright has not expired but which are no longer available - violate the constitution. ([http:// cyberlaw.stanford.edu/case/kahle-v-gonzales](http://cyberlaw.stanford.edu/case/kahle-v-gonzales))

This suit was dismissed by the Ninth Circuit, as was a similar case in the D.C. Circuit Court, *Luck’s Music v. Ashcroft*.

A third suit addressed the extension of copyright term as well as first amendment issues from another perspective. It focuses on another “enclosing” copyright issue, that of restoring copyright protections for works, in this case foreign works, that had already entered the public domain. In the words of the original complaint:

This is an action to challenge the constitutionality of Congress’s attempt to remove and radically deplete the supply of literary and artistic works from the public domain...Congress’s dramatic expansion of the term of copyright [in the CTEA] has been accompanied by an even more radical depletion of works from the public domain. On December 8, 1993, Congress amended the Copyright Act to recognize for the first time in the history of our copyright law a general provision that purports to “restore” copyrights – retroactively – in numerous works that heretofore had indisputably been in the public domain for failure to

satisfy the requirements of the Copyright Act (*Golan v. Ashcroft*, now *Golan v. Holder* 2012).

Although this suit was not dismissed, and, in fact, the 10th Circuit court held that, indeed, Congress's removal of works from the public domain that were already part of the public domain reached the Supreme Court's definition of changing the "traditional contour of copyright" and remanded the case to the district court for trial. The government, defendant in the trial, requested an *en banc* hearing by the entire 10th Circuit bench. That request was denied. In April, 2009, the District Court for the District of Colorado granted a motion for summary judgment in *Golan v. Holder*, accepting the change in the "traditional contour of copyright" argument. In the words of the plaintiff's attorneys: "It is the first time a court has held any part of the Copyright Act violates the First Amendment and the first time any court has placed specific constitutional limits on the government's ability to erode the public domain." (Falzone 2009) That decision was later reversed by the Tenth Circuit Court of Appeals. The plaintiffs appealed to the Supreme Court which affirmed the Tenth Circuit's decision and held that the government did not exceed its authority in removing the formally public domain works from the public domain

Like the legislative initiatives mentioned above, court challenges to extensions of copyright have universally failed. Prospects for relief through Congress or the courts, at the moment, do not seem bright.

2.3.3. Legally Re-Interpret

Underlying any legal statute concerning intellectual property, and thus copyright, is a set of assumptions about what “property” actually is. Laws such as the DMCA have emerged because the forms of property have changed in the digital age, while the conceptualization of the nature of property has not. Consumption, excludability, costs of replication, and other characteristics of physical property may not apply in the same way to intellectual property as to physical property, yet recent legislation and court decisions seems to assume they do.

In the last two decades, and particularly in the past decade, some scholars have argued that intellectual property and physical property such as land are not the same thing and that, in fact, the set of assumptions underlying laws governing intellectual property in a digital environment should not be based on the analogy of physical property but rather on some other model more reflective of the nature of intellectual property itself. As Wesley Hohfeld has famously pointed out, intellectual property claims are claims between people (Hohfeld 1978), not, as earlier legal commentators described, claims of people on something inanimate but tangible such as land.

Why this upwelling of legal theory with respect to intellectual property now? Simply put, the need did not exist as urgently before.

Until 1976, using the model of physical property as the basis for copyright law worked reasonably well. “Excludability” had to be claimed through copyright registration, which only a minority of creators sought to assert, and that excludability was tempered by first sale and fair use rights of users of the

intellectual property. Economically, there were significant burdens encountered in large scale copyright violation. Any type of large scale violation of copyright required a significant investment, for example, in printing press equipment or video and film duplication equipment. In this environment, the analogy to physical property, despite the clear differences in intellectual property (e.g., it is non-rivalrous), worked well enough.

Then came digital and the Internet. The economic burdens of making perfect copies and distributing them widely almost completely disappeared. At the same time, the technology to enable creators to exclude potential users from the use of their works – supported by civil and criminal law – became widely available. Now the differences between physical property and intellectual property were thrown into sharp contrast, and legal and economic theorists began to respond.

We think of information as property; law and economic structures, we argue, make it so. But this should not be the end of our inquiry. If we believe information is property, we must ask: What *kind* of property is information (Heverly 2004)?

Recent theorists have approached an answer to this question in a variety of ways. Heverly, for example, concludes that “information is not a private property regime: it is a semicommons” which, in his analysis, reflects the “dynamic relationship and interdependence of private and common property interests.” P2P file sharing, for example, represents such an interdependence. On the one hand, P2P sharing of music may have a negative economic impact on a copyright owner by reducing some potential sales of a piece of music; on the other hand,

the exposure and “word of mouth” available through P2P file sharing has a positive economic impact and increases sales and thus income for the same owner (Heverly 2004).

In fact, some music companies are actually using P2P file sharing activity statistics to promote future “hit songs” to radio stations. They are doing this promotion through third parties in order not to dilute their claims of harm due to copyright infringement, since music companies are simultaneously suing those who distribute copyrighted music through P2P networks.⁵ Leaving aside the contradiction involved in these apparently conflicting activities, this example is precisely the type of interdependence that Heverly posits as a characteristic of a semicommons model of property.

Jacqueline Lipton asserts that there is nothing wrong with viewing information as property in the traditional sense, as long as property rights and obligations are viewed in a holistic manner. Problems arise when there is an imbalance in the rights and obligations of property owners: “the problem can be re-cast in terms of the ‘absolutism’ of information property rights...” (Lipton 2004) Lipton argues that even physical property rights are not absolute, and neither should information property rights be:

Traditional property theory has always addressed the balance between private rights and public interests in property. The Hohfeldian “bundle of rights” idea of property, for example, contemplates not only rights in property, but also obligations owed to society in respect of property (such as the obligation to maintain premises in

good repair). The Lockean property concept also contemplates obligations owed by a property owner to society, such as the obligation not to waste resources, the obligation to leave “as much and as good” in the common for the use of others, and the obligation not to harm others through an appropriation of resources from the common. It is possible to create information age equivalents to these public obligations. Information property owners could be made liable for legal and financial burdens inherent in facilitating identified public interests in information. Some relevant public interests might include privacy rights in personal information, public access and use rights in scientific/ technological/ educational information, moral rights in “information works”, and/ or cultural rights in information (Lipton 2004).

Her point is that “where a government has created, or supported the creation of, private rights in information, it should be prepared to create and support concurrent public duties” (Lipton 2004).

Lipton shares a conclusion, if not the process of arriving at that conclusion, with Mark Lemley. He quotes with approval the view of the Supreme Court of Canada in *Compo Co. Ltd. V. Blue Crest Music Inc.*:

copyright law is neither tort law nor property law in classification, but is statutory law. It neither cuts across existing rights in property or conduct nor falls in between

rights and obligations heretofore existing in the common law. Copyright legislation simply creates rights and obligations upon the terms and in the circumstances set out in the statute (quoted in Lemley 2004).

In short, Lemley argues that intellectual property is *sui generis* and needs to be envisioned as such when crafting legislation to define appropriate economic rights, characteristics, and obligations rather than to use terms of “inapposite economic analysis borrowed from the very different case of land.”

All of these legal scholars find the root of the enclosure problem with respect to information to lie in the legal assumptions underlying the legislative and judicial analysis of the nature of intellectual property. They, as well as others (e.g., Breyer 1970, McCarty 2002, Lunney 1996, Pessach 2008, Sohn 2007, Sprigman 2004, Samuelson 2007, Boldrin and Levine 2002, Parchomovsky & Weiser 2010) propose alternative legal and economic analyses which, in their views, would go a long way toward reducing or eliminating at least some of the legal aspects of the enclosure of the information commons.

Fair Use has traditionally been the balancing mechanism in the copyright social contract. However, in the eyes of some scholars, the advent of works in digital form along with technological DRM protections have weighted that balance heavily on the side of rightsholders to the detriment of Fair Uses on the part of consumers.

The more technology reflects only one set of interests, however, the more it departs from the law, which conceptualizes copyright as a balancing of interests, with the ultimate goal of

fostering both creative expression and broad public availability of creative works. The result has been a perverse scenario nowhere commanded by the Copyright Act or the DMCA, in which technological measures have been allowed to override the fair use doctrine (Armstrong 2006).

This is not simply a theoretical problem, nor one confined to the United States. Lynne Brindley, CEO of the British Library stated in 2007 that:

It seems to me, as CEO of the British Library and therefore representing the researcher in part, that the balance that is referred to here—between private rights and public domain, between free competition and monopoly rights—is not working; it is being undermined by a number of things from our perspective including:

- A restrictive use of new technology (Digital Rights Management)
- Poor or outmoded legislation (i.e. too complex, increasing durations and harmonising durations ever upward etc)
- The public interest aspects of copyright being undermined and made irrelevant by private contract (Brindley 2007).

The issue has become widespread enough to involve the policy making bodies of some of the largest scholarly organizations in the world. For example, The Public Policy Committee of the ACM in its “USACM Policy Recommendations on Digital Rights Management” recommended that:

Because lawful use (including fair use) of copyrighted works is in the public's best interest, a person wishing to make lawful use of copyrighted material should not be prevented from doing so. As such, DRM systems should be mechanisms for reinforcing existing legal constraints on behavior (arising from copyright law or by reasonable contract), not as mechanisms for creating new legal constraints. Appropriate technical and/or legal safeguards should be in place to preserve lawful uses in cases where DRM systems cannot distinguish lawful uses from infringing uses (Public Policy Committee of the ACM 2006).

Not surprisingly, legal scholars have begun to re-think approaches to Fair Use in the digital age as well suggesting approaches which, in their views, would help to reestablish balance between rightsholders and users. Armstrong (2008) proposes a regime of what he refers to as "Fair Circumvention" of DRM technologies. Reichman, Dinwoodie and Samuelson (2007) propose a "Reverse Notice and Takedown Regime" under the DMCA in which those who would assert a claim to legally circumvent DRM for Fair Use purposes notify rightsholders they intended to take such circumvention steps, and rightsholders would have 14 days to object. The details of these proposals are not the issue here. What is of import is the effort to reinterpret law to reflect the changes in the social contract that digital technologies have made possible.

Another stream of quasi-legal thought focuses not on definitions of intellectual property nor on the empirical economic, political, or legal validity of

arguments in support of copyright extension. Rather these arguments assert that access to information is a right, based upon ethical principles as well as charters and statements of rights such as those authored by treaty organizations such as the Universal Declaration of Human Rights (Articles 19 and 27) as well as numerous non-governmental organizations such as the Library Bill of Rights of the American Library Association. Drawing on these and similar national and international declarations, some scholars have argued that “the right to access is not merely a liberty right but also a welfare right. That is, individuals’ information rights place duties on governments to provide access to information” (Mathiesen 2008).

At this point in time, these legal and moral speculations and theories remain speculations only and have to date had no real impact on access to information. However, they serve to provide a counterbalance, albeit a weak one at present, to the ongoing efforts of copyright owners to assert greater and greater control over copyrighted information in a culture that is increasingly digital.

2.3.4. Create Alternatives

In the absence of legislative or legal remedies, some have sought to leverage existing copyright law to realize goals of more open access that legislative proposals and law suits have not so far been able to accomplish.

This type of response encourages creators to forego some rights available under copyright law while retaining others. The desired effect is to widen the amount of material available in the information commons, if not in the public domain *per se*.

2.3.4.1. GNU General Public License. There is ample precedent for this tactic.

Free and Open source software has been released for over two decades under the GNU General Public License (GPL) or one of many “open source” variants. This class of licenses uses copyright law to license the use of copyrighted works under much less restrictive terms than exist under normal copyright conditions. So, for example, a work licensed under the GPL mandates that no charge can be made for the work itself (although charges for duplicating or distributing copies can be levied); that users are free to copy or modify the work as they see fit but that if any such modifications are made to the work, those modifications also must be made available under the same licensing terms as the original work. (St. Laurent 2004)

Creators use the GPL and its many derivatives and variants, such as the Berkeley Software Distribution (BSD) license, mostly in licensing free or open source software. However, similar licensing approaches can also apply to other types of copyrighted works such as text or music or photographs or motion pictures or datasets. The GNU Free Documentation License, for example, was the license underlying the text on Wikipedia for many years, one of the most popular sites on the World Wide Web, although Wikipedia has now brought its licensing terms into compliance with Creative Commons licenses.

2.3.4.2. The Street Performer Protocol. While many open source software developers are contributing time and expertise on a voluntary basis to particular open source software projects, many companies are supporting such projects by committing paid staff time to open source software development, either out of

altruism or because of possible economic advantage. This enables software developers to both “make a living” and contribute to open source software projects at the same time.

Few creative endeavors, however, attract this kind of economic support from industry. A possible solution that has been proposed, and actually implemented on a limited scale, is some version of The Street Performer Protocol (Kelsey and Schneier 1999). The basic idea is simple: today it would be called crowd-funding. Basically, a creator posts a notice that he or she will produce a particular work if those interested in viewing (or listening to, etc.) the work contribute a specific amount of funding. For example, an author may offer to produce the next chapter or the next book in a series if he or she is promised some set amount of money. Once that amount is reached, the work is produced and released to the public, both those who contributed financially and those who did not. In many cases, the work is released into the public domain or under a “some rights reserved” license and is openly available digitally.

Unlikely as this scheme sounds in a society that is organized around proprietary publishing and copyright protocols, it has actually had some success, although generally, to date, with smaller works such as songs or performances of public domain music performances. Sites like Musopen (www.musopen.com) serve as intermediaries between performers who are willing to perform works and release the performances into the public domain, and music appreciators who wish to have performances of those works available without the limitations of copyright.

Thus far, this type of arrangement has had limited application but it does offer an alternative mechanism for the production of creative goods to generate economic rewards for creators while at the same time making their creative works available under less than full copyright restrictions.

Another initiative which creates the same result using a different mechanism is the Creative Commons.

2.3.4.3. Creative Commons. Several of those who had been involved in some of the litigation summarized above decided that, while it was necessary to continue to challenge in court the validity of laws limiting access, something needed to be done at once to create alternatives to the closing off of the commons they felt was underway, and the Creative Commons was born.

The Creative Commons extends and broadens the “some rights reserved” approach of the GPL to licenses that creators can apply to a wide variety of creative works. The same digital technology that has made it possible for copyright owners to impose restrictive licenses on works in digital form also allows copyright owners to offer much less restrictive licenses for which users do not have to seek prior permission to use, as long as users adhere to the conditions set out in the license.

Typically, those conditions are much more liberal than those that obtain under copyright law *per se*. For example, Creative Commons offers a set of conditions that creators may choose to apply one or more of to their works to create a license. These are the choices creators are offered at the Creative Commons web site (www.creativecommons.org):

- Attribution. You let others copy, distribute, display, and perform your copyrighted work — and derivative works based upon it — but only if they give you credit.
- Noncommercial. You let others copy, distribute, display, and perform your work — and derivative works based upon it — but for noncommercial purposes only.
- No Derivative Works. You let others copy, distribute, display, and perform only.
- Share Alike. You allow others to distribute derivative works only under a license identical to the license that governs your work (<http://creativecommons.org/about/licenses>).

Creative Commons takes whatever conditions the creator indicates she wishes to attach to her work, and creates a legal license that the creator attaches to the work. The license comes in three forms: (1) human readable (a general description of the license terms in common language), (2) lawyer readable (a legal language license), and (3) machine readable. The creator indicates that the work is licensed under a Creative Commons license, and provides a link to the Creative Commons website where the specifics of the license are laid out for any potential user to view. As long as the user conforms to those conditions of use, there is no need to track down the copyright owner and obtain specific permission to use the work.

While these Creative Commons licenses do expand access to information in a commons spirit, the works are licensed under copyright and the licenses chosen draw their force and enforceability from copyright law. None of these licenses

has yet had its validity fully tested in court in the U.S. although there are instances of courts in other countries upholding the validity of the Creative Commons licenses.

Some creators are uncomfortable with having their work under copyright for 70 years after their deaths. For these creators, the Creative Commons also offers a “Founder’s Copyright” option. This option limits a creator’s claim to copyright to 14 years, the original grant of copyright in the U.S., after which time the work enters the public domain. Creators may also choose to simply affirmatively donate their work to the public domain immediately, and Creative Commons provides a mechanism for doing that as well. U.S. law makes no specific provision for this type of dedication so the Creative Commons dedication is as close as a creator can come. Before 1976, a work entered the public domain unless copyright was registered. Now, as noted above, it is necessary to specifically disavow copyright ownership for a work to be considered in the public domain.

In the years since Creative Commons licenses have become available, creators have applied Creative Commons licenses to an estimated 880 million works as of mid 2014, and the rate of use has been growing steadily. While Creative Commons supporters do not pretend that this is more than a small percentage of created works on the Internet, they do assert that it is important to have a legal channel available for those who wish to contribute to the expansion of the information commons, even if not to the public domain itself.

Alternative licensing schemes such as those employed by Creative Commons or the Open Source software movement do create a mildly competing economic model to traditional markets in copyrighted material. Creators under these

alternative licensing systems do not generally attempt to capture all value of their work but choose instead to reserve only some value for themselves. Some universities are incorporating Creative Commons licenses into their institutional structures. Stanford University, for example, no longer requires that theses and dissertations be microfilmed. Now they are simply made available electronically under Creative Commons licenses.

Open access publishing initiatives go even further and actually create an alternative model of academic publishing that competes directly in the market for academic scholarship.

2.3.4.4. Open Access Publishing. Scientific progress depends on scientists having wide-ranging access to scientific information. The same confluence of forces that has adversely affected the information commons in general has adversely affected the scientific commons, according to many in the scientific community and the communities of information professionals who serve them.

While there are over 24,000 scientific journals currently published by 2000+ publishers (ÓhAnluain 2004), fewer than half a dozen large publishers own or control the distribution of a large majority of those journals, including a majority of the intellectually most important ones. These publishers are in a quasi-monopolistic position and have been raising prices in excess of increases in the rate of inflation for two decades. During the past decade, publishers also have increasingly migrated their publications to digital form, in many cases abandoning paper publishing altogether.

Once their products are in digital form, publishers are in a position to impose technologically enforceable licensing controls, and most have done so. One result of this technologically enforceable quasi-monopolistic position is entirely predictable under capitalistic economic theory. Publishers, unfettered by competition, have bundled many titles into packages in a “take it or leave it” fashion, and have unilaterally set price points to maximize profits. The strategy has worked: the industry reported profit margins of 40% in the middle of the first decade of the 21st Century (ÓhAnluain 2004) and while profits may have declined somewhat since, they are still in very healthy double digit territory.

Academic publishers pay nothing for the articles they publish. Scholars who submit articles for publication are typically university faculty who are being paid to do research and for whom publication is part of the research process. Publishers pay nothing for the peer reviewers for the same reason. This “free labor,” combined with an increasingly non-print distribution environment, reduces costs dramatically. When combined with a near monopolistic pricing ability, these advantages result in enviable profit margins.

They have also resulted in an increasing tide of customer resentment. Scholarly libraries have had to continually cut back on journal purchases and/or reduce monograph purchases in order to attempt to keep up with rising journal prices. In many libraries, journal purchases now make up two-thirds or more of acquisition costs, with only one third going for books and other materials.

Since scholars typically sign over copyright to publishers, some scholars have found themselves in the ironic position of not being able to legally provide copies of articles they have authored, and which they provided to publishers for free, to

their students because their libraries can no longer afford to purchase the journals the articles were published in.

In this environment, libraries and librarians began to react, as did a host of non-governmental and professional organizations. One of the clearest statements of their view of the recent situation with respect to academic publishing is included in this description of SPARC:

Scholarly Publishing and Academic Resources Coalition is an alliance of academic and research libraries and organizations working to correct market dysfunctions in the scholarly publishing system...Its strategies expand competition and support open access to address the high and rising cost of scholarly journals, especially in science, technology, and medicines—a trend which inhibits the advancement of scholarship (SPARC 2006).

SPARC, as well as many other organizations, encourages the development of open access journals, publications which make their articles available to the public at no cost to the user, and which typically allow the user to make copies in digital form, and often confer a wider set of usage rights. These efforts have had some notable success.

The Directory of Open Access Journals (DOAJ) lists over 10,000 peer-reviewed open access journals containing nearly 1.2 million articles as of February of 2015. (www.doaj.org) Faculty Senates and other policy setting bodies in educational institutions in this country and abroad, including major institutions such as MIT, Harvard, and Stanford as well as over 30 other U.S.

universities and colleges, have voted to make open access mandatory for their faculty members. However, open access publishing still accounts for a small proportion of the articles published in scientific, technical, and medical journals each year, to say nothing of journals in other fields.

Funders as well as government agencies are beginning to take notice of the effect of limitations on access to information on scholarship and learning. A discussion of the impact on access by these organizations is in Chapter 3.

There are still many important obstacles for open access publishing to overcome to be a full-fledged market alternative to commercial publishing, including building sustainable economic models and changing the culture of academia to value open access and traditional publication credits equally when considering tenure and promotions. Nonetheless, open access scholarly publishing is already having an effect on the marketplace and, through market mechanisms, has already begun to expand the information commons.

2.3.4.5. Open Access to Data

While much of the effort in open access publishing is focused on finished scholarly publications, there also has been a good deal of activity designed to make data underlying those publications available under open access principles. Scores of new data sharing initiatives in scientific disciplines have sprung up in recent years as data sharing, mining, and exploration becomes more and more critical for the conduct of science, especially “big science” fields such as astronomy and genetic sciences.

In the following chapter, we review those efforts with special focus on spatially-referenced data to describe a possible place that a Commons of Geographic Data could occupy within this larger scientific data context.

CHAPTER 3

ACCESS TO SCIENTIFIC DATA IN THE 21ST CENTURY: RATIONALE AND ILLUSTRATIVE USAGE RIGHTS REVIEW

3.1. Introduction

Data has been, and remains, the lifeblood of science. For nearly 400 years, the scientific method has depended on access to data to move knowledge and society forward. That tradition stalled to some degree in the second half of the 20th century and the beginning of the 21st. For a variety of economic, legal, and, to some extent, professional reasons, access to scientific data today is not nearly as open as many wish. That situation has been changing in recent years due to a variety of societal and scientific forces, yet obstacles to open access to scientific data still exist, especially in the area of clearly delineated legal rights and restrictions.

This chapter reviews some of the forces pushing toward more open access to scientific data in the 21st century. The focus is primarily, though not exclusively, on publicly funded, geospatially-related data in a U.S. context although, in today's connected world, data access often transcends political borders, especially in disciplinary contexts. This examination looks at usage policies of a selection of data repositories that are attempting to make scientific data more accessible to determine whether usage policies are clearly understandable and consistent among repositories.

3.2. The Role of Data in 21st Century Science

More than one scientist has used the metaphor of “drinking from a fire hose” to describe the huge amount of scientific data already being generated by large scale data collectors. That “hose” will only get larger as huge data generators such as the Large Synoptic Survey Telescope and the Large Hadron Collider at CERN collect more and more data. Yet “Small Science” projects are an even more important factor in the exponential growth of scientific data generated today, possibly generating two to three times as much data as “Big Science” (Carlson, 2006).

Wireless sensors, increased computing power, higher bandwidth communication, and other increasingly affordable technologies, to say nothing of the increase in the number of researchers around the world, are giving birth to data streams unthinkable even a decade ago. Data mining and analysis are increasingly important in 21st century scientific discovery, so much so that one pop-science observer penned an article entitled “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (Anderson, 2008).

While this may seem an extreme characterization, data mining, database analysis, and other data manipulation tools and processes are now central to the enterprise of science and to new discoveries. Some researchers are even developing algorithmic processes for machine identification of natural laws from data sets without any attempt to “teach” the machines before the analysis process begins (Anthes, 2009). While this may not signal the end of theory as Anderson postulated, it certainly adds a new method to scientific discovery.

In looking at a specific subset of scientific data, geospatial data, Lance McKee of the Open Geospatial Consortium has listed “Seventeen reasons why geospatial research data should be published online using OGC standard interfaces and ISO standard metadata.” Among those reasons were an assertion, based on an analog to network theory first popularly stated in Metcalf’s law, that “The value of data increases with the number of potential users” and an observation that “Data are not efficiently discovered through literature searches” (McKee, 2010).

In the U.S., the National Science Foundation has funded the DataNet Federation Consortium, one among an increasing number of efforts to create an infrastructure that will maximize the utility of data to scientists and researchers. In describing that effort, Stan Ahalt, one of the team members working on the project, asserted that “Data is the currency of the knowledge economy... [By building infrastructure] We’ll be more efficient at producing new science, new innovation and new innovation knowledge” (Tuutti 2011).

3.3. Reasons for Calls for Open Access to Scientific Data

Over the past fifteen years, there has been an increasing number of position papers and studies calling for open access to scientific data from governments, professional and academic organizations, citizen groups, and industry. The rationale driving these calls range from adhering to the traditional mores of science to stimulating economic growth to asserting access to scientific data should be considered a basic human right.

Governments and government organizations, e.g., The National Science Foundation, the National Research Council in the U.S., the European Commission and the Royal Society in Europe, have called for better access to scientific data as a means to spur innovation and economic growth because they realize that data generated by governments and made freely available for re-use can have a significant impact on economic activity. In the U.S., for example, at least 500 companies have been identified as building new businesses on freely available data generated by the U.S. Federal Government (GovLab, 2014). One of those companies began in 2004 using openly available NOAA data and sold for a billion dollars a decade later (Kash, 2014).

While the economic benefits of open access are clearly important, in this review we focus on the scientific and, to a lesser extent, social rationales for open access to scientific data.

3.3.1. Traditional Functions: Experiment Replication and Validation

Traditional science often involves replication of research to prove or disprove results, as well as testing reported outcomes using alternate approaches and experiments. In many cases, such as with data gathered on expensive expeditions or with time-series data, access to the original, non-duplicatable data is essential for the conduct of science. In an age when data are increasingly the starting point for discovery, access to data becomes even more essential for carrying out the traditional process of science.

To enable access, storage and retrieval are essential: so is knowing what can be done with the data once they are discovered. Confusion over intellectual

property rights, or outright refusal to provide access to data, is more common in science than many imagine. In a 2006 AAAS survey of academic and industry bioscience researchers, 35% of academic and 76% of industrial researchers said that their research had been adversely affected by intellectual property restrictions of one type or another. The same survey indicated that even obtaining publicly funded data often presented difficulties. Twenty-four percent of respondents who indicated they had tried to obtain data from publicly funded sources reported difficulty in obtaining such data, and this was especially true in the fields of engineering, math, and computer science. Seventy percent of those who had difficulty obtaining data reported it had “some negative effects” on their research, and 10% experienced “serious negative effect.” Perhaps even more distressing, 16% of those denied access to data from publicly funded sources were denied access to data for which results had already been published, and 44% received no reason for the denial of access (Agres, 2006).

Reports such as this one have been one impetus for the introduction of legislation in the U.S. that would make published articles in peer reviewed journals based on research funded in whole or in part by the federal government freely available after an embargo period. The Federal Research Public Access Act of 2010 was one early example. The Fair Access to Science and Technology Research Act of 2013 introduced in the 113th Congress (2013-2014) is the most recent example. This bill would make journal articles freely available six months after publication.

The Frontiers in Innovation, Research, Science, and Technology Act of 2014 (FIRST Act) would extend that hold period to 24 months with a possible additional 12 month embargo, a bill more to the liking of publishers of scientific journals. Interestingly, the FIRST bill provides that, unlike the published article itself which may be embargoed for 24 months, “in the case of data used to support the findings and conclusions of such article, not later than 60 days after the article is published in a peer-reviewed publication.” Journal publishers widely supported the Research Works Act (HR 3699 in 112th Congress), which would have prohibited open access mandates altogether.

None of these bills have passed in the Congress. However, a provision in the Consolidated Appropriations Act of 2014 requires federal agencies in Labor, Health and Human Services, and Education with research budgets of over \$100 million to provide public access within 12 months of publication in a peer-reviewed journal to research resulting from projects they fund. While these requirements do not specifically refer to data *per se*, an increasing number of publishers are endeavoring to include data as part of the publication process.

For example, publishers such as The International Association of Scientific, Technical and Medical Publishers; The Association of Learned and Professional Society Publishers; the Public Library of Science as well as individual journals, e.g., *Nature*, *The American Naturalist*, *Evolution*, the *Journal of Evolutionary Biology*, *Molecular Ecology*, *Heredity*, have all established policies requiring that data that are the basis of articles must be made publicly accessible as part of the publication process.

Connecting underlying data sets to articles in which they appear is not a trivial undertaking. Organizations such as NISO/NFAIS (2013) in the U.S. and the Digital Curation Centre in the UK (Ball & Duke, 2011) have issued standards for citing and connecting data sets to the articles in which they appeared so that the data is findable and permanently linked to published journal articles.

The National Science Foundation has made inclusion of a Data Management Plan (DMP), which indicates where data is located and how it can be shared, a required part of research grants it funds (National Science Foundation, 2011). The University of California has created a web site with “easy-to-use” tools to develop those required DMPs (University of California 2014).

In short, the traditional functioning and, in fact, the traditional *mores* of science since the Enlightenment require the ability to find data, to access them, and to be able to use them to both verify scientific claims and to extend discovery. Funding agencies and publishers alike are beginning to take steps to ensure that data discovery and access are possible.

3.3.2. Avoidance of Duplication

In an era of tight research funding and limited resources, an important reason to make scientific data available for widespread use is the wasted cost of duplication of effort, particularly when it occurs simply because researchers do not know what other work has been undertaken if data are not openly accessible. Mounting expensive expeditions to places such as Antarctica to gather what turns out to be essentially duplicative data are obvious examples of expensive and avoidable duplications of effort.

In short, reproducibility of experiments for the purpose of validation is essential to the practice of science. The practice of duplicating efforts, however, is wasteful science, and timely access to data can help to reduce such wasteful activity in an era of limited resources.

3.3.3. Access to Data as a Human Right

In the 21st century, science and technology will continue to have an enormous impact on standards of living around the world as well as on freedom and governance. This is one reason why there is an increasing interest in the claim of access to information, including scientific data, as a human right.

For some, e.g., Shaver (2009), that claim finds its source in Article 27 of the Universal Declaration of Human Rights: “Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits” (United Nations, 1948).

Others, e.g., the New York Law School/Healthcare Information for All 2015 Human Rights and Healthcare Information Project (2009), focus on a particular “right,” in this case a right to health, and this claim finds its basis in another section of the Universal Declaration of Human Rights: “Everyone has the right to a standard of living adequate for the health and well-being of himself and his family, including food, clothing, housing and medical care.”

In preparation for the UN’s 2016-2030 development agenda to succeed the UN’s Millennium Development Goals, the International Federation of Library Associations (IFLA) submitted the “Lyon Declaration on Access to Information and Development” to the UN. The Declaration includes the statement:

We, the undersigned, therefore call on Member States of the United Nations to acknowledge that access to information, and the skills to use it effectively, are required for sustainable development, and ensure that this is recognised in the post-2015 development agenda by:

a) Acknowledging the public's right to access information and data, while respecting the right to individual privacy..."

(International Federation of Library Associations, 2014).

As of December, 2014, that language is included in the UN Secretary General's Draft of Sustainable Development Goals for the next decade and a half.

Evaluating the validity of these claims that access to information is a human right is not within the scope of this review. The import here is that the assertion that access to information and data is a human right has reinforced calls for open access to scientific data from still another perspective. Some recent initiatives, while not specifically speaking to the rights claim, have seemed to support it by providing immediate open access to both reviewed papers and raw data when an emergency threatened.

A good example is one of the efforts to provide real time open access to research into the science and spread of H1N1 flu in 2009-2010 via the *PLoS Currents–Influenza* web site (Olson et al, 2011). In this case, there was an immediate emergency which this initiative responded to, and the open sharing of data became almost an imperative. Similar efforts by the general public as well as professional researchers using Google Maps or other online technology have taken place in several cases to follow the spread of a contagious disease.

None of these efforts would be possible without open access to data. Open data advocates point to examples such as these in arguing for increased access to data in the service of the health and well being, both physical and economic, of all people, often pointing to international agreements such as the UN Declaration on Human Rights as a justification.

3.3.4. Data Preservation and Archiving

Today, a tremendous amount of scientific data is “born digital,” and that fact is a source of much unease in the scientific and public policy communities. A huge amount of digital data is essentially “endangered data” and, in many cases, once it is gone, it can never be replaced (Murillo, 2014).

Examples of new discoveries being made based on existing data that the original authors had no idea about are common in scientific history. “Many classic results in science have come from the analysis of existing knowledge already available in the open literature” (Murray-Rust, 2007). With the “data deluge” today, that is likely to be even more true as machine algorithms mine ever expanding data sets in ways and at speeds that no human can match. As one researcher responding to a European Union survey on data preservation put it: “The most important reasons for preservation are the ones we do not see now” (van der Hoeven et al, 2010).

Agreement on the need for the preservation of digital data is widespread. In the U.S., the Committee on Science, Engineering and Public Policy of the National Academies of Science put the rationale for preservation very simply: “Research data should be retained to serve future uses. Data that may have

long-term value should be documented, referenced, and indexed so that others can find and use them accurately and appropriately... In some research areas, accessible databases have become essential parts of the research infrastructure, comparable to laboratories, research facilities, and computing devices and networks" (2009). This type of thinking is mirrored in reports or position papers or grant funding requirements by the National Science Foundation (2006, 2010), by the European Commission (2013), and NSF/Jisc (Arms & Larson, 2007).

While the motivation and justification for effective archiving of scientific data are widely acknowledged to be valid, what is actually happening on the ground, especially in "Small Science," often fails to capture data for archiving and re-use. In some disciplines, the estimate is that as much as 80% of data developed by individual researchers or small teams is not captured in a public way and is often simply lost over time (Murray-Rust, 2007). The National Science Board (2005) has noted that at the level of what it refers to as *Research Collections* "Authors are individual investigators and investigator teams. *Research collections* are usually maintained to serve immediate group participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards."

Kansa and Bissell (2010) have proposed a web syndication approach for sharing primary data in "Small Science." This approach, if implemented by researchers, would make distribution of data sets more widespread. Yet this approach does not specifically address preservation.

In an effort to capture data from “Small Science” *Research Collections*, as well as from larger research endeavors (both *Resource Collections* and *Reference Collections*, in the National Science Board’s terminology), universities are establishing institutional repositories that can handle data as well as publications; disciplinary repositories are being established; and some publishers are setting up data repositories to house data related to articles published in their journals.

With this flurry of activity over the past decade, the questions naturally arises: What characteristics should scientific data repositories have in order to be effective in ensuring that data will be “readily available, accessible, and usable” (Arms & Larson, 2007) and can be “easily consulted and analyzed by specialists and non-specialists alike” (National Science Foundation, 2006)?

3.4. Desirable Characteristics of Data Collection and Storage Systems

Although goals and aspirations can be expressed in general terms, operational characteristics of an effective repository environment need to be more specific. A number of workshops and reports over the past decade have endeavored to outline functions that are desirable in a data storage and access system. In the U.S., examples include *Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content* (Institute of Museum and Library Services, 2003), *Licensing Geographic Data and Services* (National Research Council, 2004), and *To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering* (Association of Research Libraries, 2006).

While those sets of recommendations differ in some ways, reports share common characteristics that the authors see as important for the preservation of scientific data for use by both current and future generations of users.

Characteristics include: access; clear use conditions; findability; interoperability; evaluation capability; and the technical issues of ensuring data integrity, scalability, and life cycle management for preservation through time.

While some of these reports are focused on very large data sets, they are readily applicable to data repositories for data of any size. We briefly describe these desirable characteristics in turn, clustering related characteristics together where appropriate.

3.4.1. Access

The first step in being able to benefit from scientific data is being able to get access to it in the first place. Data that are not online, or are hidden behind paywalls or other restrictive barriers online are not readily accessible to researchers or to the public. Part of The National Science Foundation's *Cyberinfrastructure Vision for 21st Century Discovery*, for example, describes an environment in which data "are openly accessible while suitably protected" so that they may be "regularly and easily consulted and analyzed by specialists and non-specialists alike" (2006).

3.4.2. Clear Use Conditions

Accessibility by itself, as the NSF's words above suggest, does not guarantee the ability to re-use data. To be maximally useful to others, data sets

must carry with them information about how they may be used, e.g., through clear licenses. While facts *per se* are not copyrightable in many political jurisdictions around the world, including the United States, it is often difficult to tell whether an arrangement of facts is original enough to afford copyright protection and could restrict or limit entirely the uses to which data can be put. In a world of increasingly internationalized repositories, data originating outside of the U.S. may have other legal or restrictions on use, e.g., *sui generis* provisions in the EU. Absent a clear indication by those who produce data sets indicating to what uses the data sets may be put and conditions on their use, if any, the data is essentially useless to others. The uncertainty about possible consequences of misuse will deter most present and future potential users from employing the data for new purposes.

3.4.3. Findability

In a time of ever-increasing growth of scientific data, being able to find what a user is looking for in a sea of data becomes critically important. Findability depends upon being able to search for data in a consistent manner and in having that data be identified in a consistent manner over time so that they are always discoverable. Both finding particular data across time and space and then being able to access that data depend heavily on standards-based metadata. Data must also have a consistent and permanent identity and location identifier over time. Put simply, if a user cannot find data s/he is seeking, they will never get used.

3.4.4. Evaluation Capability

Science, for at least the past 400 years, has been based on peer review. Repositories or other data collection structures help to make data more valuable when those interested in the data can, if not formally review them, at least comment upon them and discuss their usefulness for particular purposes. In the case of irreproducible data (e.g., time series data, data gathered on expeditions that are not likely to be repeated, etc.), discussions about the data themselves, methods of collection, and so on are critically important. Using the data for other purposes, applying new tools in the future with which to analyze data or even re-analyzing samples collected and stored that are made visible through the metadata in repositories all benefit from having access to the comments of prior users.

3.4.5. Technical Characteristics

For data sets to be useful for future research purposes, users must have confidence in the integrity of the data set. Life cycle management will require data being transferred from one storage medium to another on a routine basis over time to ensure accessible preservation. If any corruption of that data takes place in the process, or for any other reason, the data becomes suspect, at best. Repository sponsors are naturally concerned with ensuring data integrity, and research is under way to develop standards and best practices for data handling and preservation. From developing unique hash based identities to keeping redundant copies, efforts are underway to ensure users that the data they access are an exact copy of the data that were contributed to the repository or

collection. “Lots of Copies Keeps Stuff Safe” (LOCKSS), for example, is software that not only allows institutions to keep redundant copies of information but also regularly audits files at the byte and bit level and repairs them on an ongoing basis (LOCKSS, 2014).

In addition to managing data integrity over time, effective preservation of scientific data in today’s world requires scalability, the ability to grow storage and access capabilities and still operate reliably and efficiently. Computer scientists and database designers are constantly working to reduce uncertainty in system performance while dealing with exponential growth of the data to be preserved. At present, a new focus is developing on decentralized and virtual storage and access facilities, often run by large commercial organizations such as Google and Amazon. “Cloud-based” storage offers institutions, especially smaller ones, the opportunity to have both scalable repositories and redundancy without building physical infrastructure themselves.

And wherever data reside, interoperability is a key challenge. Data file structures and layout often differ from one data set or data base platform to another. Metadata are often inconsistent when they exist at all. Searching disparately formatted data sets is a huge challenge. Designing ways to enable a user to search across file structures and types of scientific data and come up with comprehensive and accurate results is the subject of ongoing research. While existing data sets may never be fully interoperable, efforts such as DataNet in the U.S. are working to build structures that may help future data interoperate more effectively.

3.5. A Brief Overview Of Recent Initiatives To Provide Open Access To Scientific Data

The calls for access to scientific information are being heard and acted upon in many quarters today. There are now hundreds of data repositories available online. Some were established and operated for a while but no longer seem to be maintained although they are still accessible, e.g., GlycomeDB or antbase. Some have merged with others in the same domain to provide more efficient operation, e.g., ORegAnno. Many others are still current and vibrant.

3.5.1. Open Access Data Repository Growth

In this ever changing environment, finding online data repositories is becoming increasingly difficult unless the URL is already known. Not surprisingly, this challenge has given rise to the creation of a number of data repository cataloguing and search sites. These sites provide lists of repositories and offer various ways to search for particular types of data.

The Open Access Directory (http://oad.simmons.edu/oadwiki/Data_repositories), for example, lists over a hundred directories or repositories in over a dozen different disciplines in which there is at least some open access to data. DataBib (<http://databib.org>) lists almost a thousand research data sites as of this writing, as does re3data (www.r3data.org). In an effort to provide a more centralized access point and more complete search service for data repositories throughout the world, DataBib and re3data have agreed to merge their catalogs by the end of 2015.

While these catalog sites are operated by organizations, some sites that offer data search and access capabilities are maintained by individuals. One very useful such directory of geographic data sets, Freegisdata (<http://freegisdata.rtwilson.com>), includes a list of over 300 sources of “free as in free beer” geographic data sets sorted by the type of data they contain although information varies as to whether particular repositories are also “free as in free speech,” i.e., what usage rights are. Governments, too, are endeavoring to provide access points to data repositories they provide. Some U.S. examples are discussed in the following section.

Even a cursory look at repository sites confirms that science data repositories include a wide range of capabilities and coverage, ranging from small prototypes to sites containing access to great stores of data from, for example, space probes (e.g., <http://nssdc.gsfc.nasa.gov/>), automated astronomical telescopes (e.g., <http://tdc-www.harvard.edu/>), or the Large Hadron Collider (<http://opendata.cern.ch/>).

Few of these sites are interoperable in terms of shared metadata schema or data formatting; few have anything resembling a life cycle management plan; few have a commenting or evaluation capability. Still, their existence demonstrates that there is a widening realization that providing access to, and preservation of, scientific data is a valuable and worthy endeavor. The challenge is to make generated data more widely available. Such a goal brings with it many challenges, especially with Big Data, and organizations are currently trying to clearly identify the spectrum of challenges involved and ways to deal with them (e.g., CODATA/ICSU, 2014).

It is not surprising that data that require a huge financial investment to generate, such as astronomical data from the Hubble Space Telescope, are often funded by government bodies. In the U.S. and in many other countries such data are made freely available for anyone's use although that is not the case in every jurisdiction worldwide. In large multinational efforts such as the Global Earth Observation System of Systems (GEOSS), for example, which includes 84 countries and 54 additional Participating Organizations, settling on common usage licenses for data made available through www.geoportal.org by many different countries and agencies remains a significant challenge (Onsrud et al, 2010).

3.5.2. Access To U.S. Government Generated Data

The U.S. federal government collects and generates enormous amounts of publicly funded data useful to science as well as to industry and the general public. In recent years, the federal government has been attempting to make the data it collects available for research and for simple daily use by anyone. The same is true to different degrees for governments in other parts of the world.

In the U.S., the recently launched Data.gov web site is one example. It provides access to data collected by 18 federal agencies, currently containing well over 100,000 data sets. Because the U.S. government cannot hold copyright on materials it generates (U.S. Code, Title 17, S.105), there is no claim of copyright on any of the data sets, even if they might qualify for copyright protection if generated by non-federal sources.

The U.S. federal government makes both data and tools available for use by anyone who wishes to access them. Sites such as The National Map (<http://nationalmap.gov>) provide a starting point for geographic information. The U.S. also makes life science data of various kinds available through the National Institutes of Health for both professional researchers (e.g., PubChem: <http://pubchem.ncbi.nlm.nih.gov/>) and for lay users (e.g., MedLine Plus: <http://www.nlm.nih.gov/medlineplus/>); geologic data through the U.S.G.S.: <http://www.usgs.gov/>); and so on.

While the importance of access to data is mirrored at the state and local level in the U.S., access to that data and re-use conditions are much more mixed than on the federal level.

3.5.3. Access To Data In The U.S. Generated By Non-Federal Government Bodies

State and local governments in the U.S. may hold copyright to datasets that they generate that qualify for copyright protection. Some states and some local governmental bodies are making conscious efforts to make their spatially referenced data available with no or minimal conditions on its use. Maine and Montana are good examples on the state level. Both provide significant collections of spatial data available to users, in Maine through the Maine Office of GIS (<http://www.maine.gov/megis/catalog>) and in Montana through the Montana Geographic Information Clearinghouse (<http://geoinfo.msl.mt.gov>).

MetroGIS (<http://metrogis.org>) in the Minneapolis/St. Paul area is a good example on a local/regional level.

Some states, and particularly local government bodies, view their data as a source of income and resist efforts to make it accessible at no cost and under minimal reuse restrictions. This is particularly true for spatially-referenced deed, tax, and other information associated with real estate and real property. Even in states with strong Freedom of Information laws, some municipal and county governments seek to hold onto control over access to data, especially when it is in electronic form, out of concern that the income potential for the government body will be reduced if other entities get access and then make the information available at low or no cost (e.g., the case of Brick Township, NJ: <http://www.rcfp.org/news/2005/0712-foi-utilit.html>).

There are also other motivations for limiting access to information collected by state or local government bodies. Locations of endangered species, for example, are often not made public or exact information about locations of certain types of conservation easements granted to towns out of respect for the privacy of the donors.

Whatever the justification, access to locally generated data at the non-federal level in the U.S. is much more varied than access to data generated by the federal government.

3.5.4. Private and Corporate Initiatives

While the focus of this review is primarily on publicly funded data, it is important to note that although private companies usually view their data as

proprietary, there are cases in which they make that data available for use at no charge even though they retain ownership.

In the area of spatially-referenced data, Google Earth, Google Maps, and related services by providers including Rand McNally, Mapquest, and others offer access to various types of spatially-referenced information through both computer and mobile devices that are now a part of everyday life for many people. While widely used, including in academic and government contexts, these services lack important features that dependable open access and archival services should include.

First and most simply, these services are proprietary, and even if a company's public goal is "Don't be evil" (as Google's is), there is not and cannot be any guarantee that policies in private companies, especially publicly traded stock companies, will not change when shareholder value demands it. Company policies and practices can change abruptly, as any of Facebook's billion users or the millions of users of Google's gmail service or even Google Maps well know. Building access to scientific information on proprietary foundations is risky as far as guaranteeing access to, and preservation of, data into the future is concerned.

In addition, even though services such as Google Earth allow contributions of spatially-referenced information from users, questions about usage rights and provenance of posted information abound, and there are no metadata standards in use for contributed information. While keyword search mechanisms have considerable power, they are simply inadequate for scientific search and retrieval purposes, and this is particularly true in the case of spatially-referenced

data. In addition to these considerations, the question of the quality of Volunteered Geographic Information (VGI) is also an unsettled one (Flanagin & Metzger 2008).

Private companies may also offer access to a subset of their tools and data for a combination of public service and quasi-promotional purposes. Often these are educational endeavors such as ESRI's ConnectEd Initiative (<http://connected.esri.com/>) which, while providing students and teachers with classroom tools, also introduce students to the company's products.

In dealing with medical data, as another example, private companies sometimes find it in their interest to make some of their data publicly accessible. When private companies do so, they often, as in the case of clinical trial data made available by some pharmaceutical companies to the Yale Open Data Access Project (<http://yoda.yale.edu>), retain proprietary ownership of their data and are free to remove them from public sight at any time.

In short, private and corporate initiatives can be welcome supplements to, but at present are unlikely to be major contributors of, openly available scientific data.

3.5.5. Non-U.S. Access Efforts

While the primary focus of this review is on U.S. policies and access efforts, in today's international environment, it is impossible to ignore the access to primarily publicly funded scientific data in other countries. Many large repositories, especially disciplinary repositories, include data originating from different countries. In some cases, those repositories have a single policy

regarding access and re-use, but in many other cases, access and re-use policies are tied to the laws in the countries from which the data originates. Countries around the world, most of which are able to hold copyright on data, have varying policies on access and re-use. An overall review of those policies is not appropriate here, but it is worth noting that many countries are making efforts to make government generated data, especially geodata, more widely open and available. Examples include UK Location (<http://location.defra.gov.uk>) in the United Kingdom, the Atlas of Canada (<http://atlas.gc.ca/site/english/index.html>), and Geoscience Australia (www.ga.gov.au), all of which provide open access to some government-generated spatially-referenced data. In the brief review below, we include some sites that include non-U.S. data and/or are non-U.S. based for illustrative purposes.

3.6. Usage Rights And Data Repositories: A Brief Review

As the discussion so far suggests and as the examples in the next section illustrate, there already exist numerous disciplinary and government run repositories, particularly those designed to provide access to collections of large scale data. In “Small Science,” the picture is much less encouraging, whether those small science data gathering efforts are university or institution based or are the results of sporadic efforts to enable individuals or small local groups with locally generated data of their own to expose them and make them available for others to use.

One absolutely critical component to the reuse of data in repositories of any scale is a clear description of usage rights and conditions for data access and re-use. In some cases, repository sites simply do not even post license information or usage conditions. In others, terms like “free” and “open” are used with a variety of meanings that are sometimes only discernible by drilling deep into the site or in some cases are not specified at all.

Data repositories are usually made up of data that, even if collected on one site, originate from many different sources and often different countries. Some repositories are “federated” in that they provide links to sites where data sets actually reside but do not collect or store data themselves. In either case, data sets may have a variety of usage rights and/or conditions attached to them, and sorting those rights and conditions can be a difficult task.

Absent a definition of terms, repository search engines or catalogs may provide information on usage rights in similar language, but whose usage rights may be very different from other sites using similar language. While there is, as yet, no universally accepted definition of “open” in the context of scientific data, there are efforts underway to create a definition that can be used generally. The Open Definition, offered under the auspices of the Open Knowledge Foundation, asserts that *“A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.”* (Open Definition, 2014)

Very few repositories specifically reference this Open Definition. One that does is Open Street Map (<http://www.openstreetmap.org>) which licenses its

data under the Open Data Commons Database License (<http://opendatacommons.org/licenses/odbl>), which in turn depends upon the Open Definition.

In reviewing the status of usage rights and conditions in the context of scientific data repositories, 40 repository sites were examined. This list includes many U.S. based sites, but because of the international nature of data today, especially data located in disciplinary repositories, some reviewed sites are based outside of the U.S. Some, such as re3data.org, are operated as collaborations of organizations located in the U.S. and in Europe. Whether accessible through U.S. government, disciplinary, or even privately operated sites in the U.S. or beyond, the great majority of open data listed below are the result of publicly funded research.

In these 40 sites, 13 different sets of usage terms and conditions for reuse of the data were identified. Summary descriptions of usage rights and conditions are listed below, followed by Table 1 identifying which usage information applied to the 40 sites. A fuller description of the sites and the conditions of use and re-use are attached in Appendix A.

The list below contains simple language descriptions of usage information based on conditions available on the listed repository sites as of December 15, 2014. The numbers are referred to in the “Usage Rights” column of Table 1 below.

1. All U.S. government sites use a similar usage message: data produced by U.S. government workers is Public Domain. However, sites may contain data, datasets, or databases provided by others that may be subject to

- copyright use restrictions. Such material will be labeled.
2. Data, where copyright restrictions are applicable, is available under a Creative Commons license.
 3. Access to the data is available to the public at no charge. The author was not able to find any information about use restrictions.
 4. Site asserts copyright in all copyrightable materials including the database itself but makes data free to use for personal, scholarly, or private research purposes. Source attribution requested or required.
 5. Data is free of charge, but some data sets may have Conditions of Use.
 6. Data is free of charge but some data sets may have Conditions of Use, and those may require user registration.
 7. Data and other material remain property of original contributing organization and should be available at no cost.
 8. License for use granted under Open Canada License—attribution required.
 9. Data available for public use with attribution.
 10. Database available under Open Database License. Any protectable content is licensed under an Open Contents License.
 11. Majority of material is Public Domain. Some data provided by others may be subject to copyright use restrictions. Such material is labeled.
 12. Data placed in the Public Domain by contributors.
 13. Data available under the Open Database License. Other material available under a Creative Commons license.

Table 1. Data repository sites with usage rights referenced to the list above.

Site Name	URL	Usage Rights
Scientific Earth Drilling Information Service - SEDIS	http://sedis.iodp.org/front_content.php	3
Data.gov	http://www.data.gov	1
PubChem	http://pubchem.ncbi.nlm.nih.gov	1
Online Mendelian Inheritance in Man (OMIM®)	https://www.ncbi.nlm.nih.gov/omim	4
Montana Geographic Information Clearinghouse	http://geoinfo.msl.mt.gov/	3

Table 1 continued

MetroGis	http://metrogis.org/	3
BOLD	http://www.barcodinglife.org	3
ChemSpider	http://www.chemspider.com/	4
Freebase	http://www.freebase.com/	2
Sage Bionetworks	http://sagebase.org/	3
uBio	http://www.ubio.org/	3
ICDNS	http://www.icdns.org/	3
ZooBank	http://www.zoobank.org/	3

OneGeology	http://www.onegeology.org	7
Gateway to Scientific Data	http://cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/gateway-scientific-data.html	8
Europeana	www.europeana.eu	2
DOE Data Explorer	http://www.osti.gov/dataexplorer/	1

NEXTBIO	http://www.nextbio.com/	10
ChemBank	http://chembank.broadinstitute.org/	3

EnvBase	http://envgen.nox.ac.uk/cgi-bin/envbase.cgi	3
LTSRF	http://ghrsst.nodc.noaa.gov/	1
ZINC	http://zinc.docking.org/index.shtml	4
ADS	http://ads.ahds.ac.uk/	4
GeoGratis	http://www.geogratings.cgdi.gc.ca	8

Table 1 (continued)

NARCIS	http://www.narcis.info/index	3
GBIF	http://www.gbif.org	12
LinkedGeoData	http://linkedgeodata.org/About	13
dbGaP	http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html	6
Open Context	http://opencontext.org/	2
RRUFF	http://rruff.info/	3
PCL Map Collection	http://www.lib.utexas.edu/maps/	11

COD	http://www.crystallography.net	12
PCOD	http://www.crystallography.net/pcod/index.html	12
Biozon	http://www.biozon.org/	3
ORegAnno [latest entry 2008]	http://www.oreganno.org/oregano/Index.jsp	3
antbase [latest entry 2009]	www.antbase.org	2
AntWeb	www.antweb.org	2

Table 1 (continued)

OpenStreetMap	http://www.openstreetmap.org/	13
TOXNET	http://toxnet.nlm.nih.gov/	1
GlycomeDB [latest entry seems to be 2102 – copyright notice is 2007]	http://www.glycome-db.org/	1
OBIS	http://www.iobis.org/home	8
ChEMBL	https://www.ebi.ac.uk/chembl	5
GeoNames	www.geonames.org	2
Dryad	datadryad.org	16
WorldWideScience.org: The Global Science Gateway	worldwidescience.org	14

National Historical Geographic Information System	www.nhgis.org	12
GlobalSoilMap.net	www.globalsoilmap.net	3
FreeGISData	http://freegisdata.rtwilson.com	6
The National Map	http://nationalmap.gov	1

As the information in this table indicates, there is a wide variety of meanings attached to the term “open” in terms of use and re-use of scientific data. In a few cases, “open” conforms to the Open Definition mentioned above. But in far more cases, there are actually conditions on re-use which, if not discovered and adhered to by subsequent users, could cause significant reputational, and/or legal or financial, risks. These “non-obvious” conditions placed on the use of data that are labeled “open” could create impediments to wider use of such data in science research.

3.7. Chapter Conclusion

There is strong, though not universal, support for open access to publicly funded scientific data among governments, the research community, business and industry, and private users. While there are many challenges to overcome to make scientific data findable, technically accessible, and to preserve them effectively through time, even if these challenges are met, there is still a very

significant question of whether and under what conditions users may re-use data in online repositories. At present, usage conditions vary widely, and a user's ability to even find what usage conditions are in effect also varies widely, even in the somewhat focused domain of spatially-related data. Absent use of specific, accepted licenses, terms like "Open" can give rise to different interpretations.

As a first step toward making scientific data really open, repositories could select from one of the currently available widely recognized and standardized data licenses that promote open access and use, such as Creative Commons licenses or Open Database Licenses. Repositories could, as some do now, make accepting the conditions of the repository's chosen license a requirement for contributing data to the repository. Users would then clearly know what they could and could not do with data found in the repository.

Having to deal with a variety of such standardized licenses, even if that variety is limited, is not ideal from a user perspective, but it is far better than having dozens of variations on usage and imprecise use of terms like "open" or "free." Ultimately, the ideal would be to have a common set of usage licenses for all repositories of scientific data to help realize the significant benefits to science and society of truly open access to, and use of, scientific data.

CHAPTER 4

**POTENTIAL CONTRIBUTOR PERSPECTIVES ON DESIRABLE
CHARACTERISTICS OF AN ONLINE DATA ENVIRONMENT FOR
SPATIALLY-REFERENCED DATA**

4.1. Introduction

Data that is related to a particular geographic location is everywhere in today's online world. Individuals and businesses use cell phone location services, Google Maps and other mapping services, and a wide range of other spatially-referenced data as part of their everyday routines.

Yet there is a potentially very valuable type of data that is not part of every day online life for one simple reason: it is not discoverable online. Small locally-generated, spatially-referenced data sets could be of great value to researchers and to the general public if they were available, discoverable, and if conditions for their use were clear. At present, that is not generally the case for such privately held data sets.

There are many efforts underway to capture and make available large scale national and international data by governments and academic or professional organizations.¹ However, small local data collections have largely been overlooked, even though they could be of use to professional researchers as well as to the general public.

There have been several recommendations to construct an online Commons of Geographic Data that would provide an environment where that data could be contributed with no special knowledge or skill or large commitment of time and

effort required on the part of contributors, yet would be “findable” by others using standards-based metadata search tools (National Research Council (U.S.) Committee on Licensing Geographic Data and Services 2004, Onsrud & Campbell 2007).

An online Commons of Geographic Data (CGD) would enable potential contributors of locally-generated, spatially-referenced data to make that data available so that others could use it. In the context of this study, spatially-referenced data means any data that refers to a specific place, which includes a large majority of data today. Some examples might include a high school class project that locates and catalogs all of the trees over fifteen feet tall in a small town; a homeowners’ association that monitors the water quality of the lake on which their property is located; a historical museum that ties its photographic images to their physical locations, a list of wheelchair accessible street crossing locations, or a weekly list of products available at a particular farmer’s market. Much of this local small data is generated and stored by private parties. It is stored on private individuals’ or local organizations’ computers and is not now publicly available online so that others might use it. It is, in effect, fully or partially “invisible.”

An online commons environment is one in which users do not have to ask for permission for using the data found there. The data owner has already granted permission, if permission for use is needed, through a “some rights reserved” license as long as the user respects any conditions put on the use of the data by the owner/ contributor. Creative Commons licenses are examples of “some rights reserved” licenses.

At present, no such Commons of Geographic Data exists for such locally generated, spatially-referenced data. If a group were contemplating the design of such a commons environment, a significant question would arise: what characteristics might potential contributors find desirable that might help motivate them to make their data available through an online CGD environment?

4.2. Potential Contributor Motivation

Any discussion of possible criteria for constructing a commons type repository for spatially-referenced data brings up the question: if such a repository were built, would people contribute to it? This specific question has not been tested to date and the focus of this chapter is not to review the literature in this area. We note, however, that there is a good deal of evidence from volunteer motivations in general, and from online volunteerism in particular, to suggest that people who own spatially-referenced data would be willing to contribute it to an online commons-type environment.

People volunteer their time, skills, and resources every day in a wide range of domains ranging from volunteering in youth oriented activities (Riemer et al. 2004), to contributing to Wikipedia (Nov 2007), to helping out as a tourist guide (Anderson and Shaw 1999), to helping predict protein structures online (Cooper et al. 2010), to contributing content and tools online (McKenzie et al. 2012) and to dozens, if not hundreds, of other activities. In short, there is an extensive literature on this subject.

Perhaps the most relevant comparison lies in the area of what has come to be called Volunteered Geographic Information (VGI) (Goodchild 2007). The explosion of effort in this area in the past few years provides compelling evidence that data owners would be likely to volunteer their data. The real question is under what circumstances contributors might be willing to contribute their data. That is the focus of this research.

4.3. Desirable Characteristics of Data Repositories

There have been a number of studies and recommendations about desirable characteristics for the preservation of data in online environments such as the *Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content* (Institute of Museum and Library Services, 2003), and *To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering* (Friedlander and Adler 2006).

Three key recommendations emerging from these and other studies are that these online environments should make it possible (1) to clearly specify usage rights, (2) to search for and discover data using standards-based metadata, and (3) to evaluate data for suitability for a user's purpose.

These may seem like common-sense ideas, and they are. We might assume that any potential data contributor to a CGD would agree with them. But that would simply be an assumption. Assumptions may be right, or they may be wrong: without empirical evidence, there is no way to judge. Research is necessary to confirm or refute these, or any, assumptions.

This study sought to empirically explore whether potential contributors to an online commons environment for locally generated, spatially referenced data found these three recommendations desirable. While not the purpose or focus of this study, the results could be useful to those who design institutional repositories at universities and colleges, as well as to others who operate or may wish to establish online data repositories for other types of locally generated small data collections.

Specifically, this research addresses the following hypothesis.

4.4. Hypothesis

Potential data contributors of locally-generated, spatially-referenced data would be willing to consider contributing their data to an online data repository with no financial compensation if such a repository included:

(a) a simple, clear licensing mechanism so that there is a way to choose which usage rights the owner is willing to pass on to users and which usage rights the owner wishes to retain, if any²;

(b) a simple process for attaching descriptions to the data. These “plain English” user descriptions would be processed by the system into standards-based metadata without requiring knowledge of metadata systems or controlled vocabulary terms on the part of the contributor;

(c) a simple post-publication peer evaluation/ commenting mechanism that would both provide feedback for contributors, and provide information on quality and suitability of use for future users.

4.5. Method

In order to test this hypothesis, we used a combination of qualitative and quantitative research procedures (Onwuegbuzie and Leech 2004, Ragin et al. 2004). Personal interviews were conducted with ten people who either had generated data of their own, or who had the authority on behalf of the groups they represented to make data generated by the group available for use outside of the group.³

To confirm or refute the findings from these qualitative interviews, we designed an online questionnaire based upon the results of the interviews, and compared results from that questionnaire with the results from the interviews.

In order to minimize bias introduced by information discussed in the interview itself, interviewees were given short pre and post-interview questionnaires to see if their opinions had changed about any of the topics discussed in the interview.

4.5.1. Interviewees and Data Types

The interviewees and/or the organizations they represented held a variety of different types of data, all of which was locally generated, and spatially-referenced in some way, and none of which was available online at the time of the interviews. The only selection criteria for an interviewee were: a willingness to consider making their data available in an online repository without any financial payment; personal ownership or legal control of that data; and a willingness to meet with a researcher in person for up to one hour.

The interviewees so chosen are not in any way a statistically representative sample of potential data contributors to an online commons environment for spatially-referenced data. The major reason for not attempting to select a statistically representative sample of potential contributors is that the number of such contributors is unknown and probably unknowable. Thus, we conducted qualitative in-depth interviews, and then used an online quantitative survey to support or refute the qualitative findings. The goal was to produce findings that would be informative, even though not “proven” in a statistical sense. The hope is that the findings would be useful for future designers of a Commons of Geographic Data type online environment, if one should be constructed.

Interviewees were selected using a “snowball” technique. Initial interviewees were suggested by people interested in data collection who were located in geographic areas accessible to the interviewer. Those who participated as interviewees recommended other potential interviewees. As chance would have it, the final group of ten interviewees turned out to be quite diverse in the types of data that they owned or controlled.

Four of the interviewees were either paid or volunteer leaders of local groups concerned with environmental and/or land use matters. Among them, these groups collected data on water, soil, and air quality; invertebrate populations; locations of threatened species; maintenance schedules for trails on preserved land; owner granted easements on private land; and other similar types of data.

One interviewee served on a town recreation committee that focused on recreational uses of water bodies in the town, and had data on resident’s recreational interests as well as on water quality in local lakes. One interviewee

was a graduate student working on a project involving ocean currents and ocean water characteristics at different depths. One high school teacher taught use of GIS software for mapping social data such as street light locations and their possible correspondence to crime statistics. One interviewee was an author of books about birding who combined the author's original data on bird sightings with state habitat maps. Another interviewee worked in an organization with an extensive collection of photographs of historical maritime objects related to specific ports. One worked with a local historical society on locating, describing, photographing, and mapping gravestones in town cemeteries.

For those interviewees working with organizations, in all cases, the organizations are non-profit, all with less than five paid staff.

Seven of the interviewees were from Maine, one from Massachusetts, one from Pennsylvania, and one from North Carolina.

4.5.2. Qualitative Data Collection Process

The purpose of these qualitative interviews was to test whether the hypothesis above would hold. All interviews were conducted from the same interview instrument by the same interviewer. The interviews were audio recorded, transcribed, and coded; and then the transcripts were checked against the voice recordings for accuracy. A summary of key points of each interview was then sent to the interviewee for correction and confirmation. None of the interviewees who responded submitted any corrections other than spelling errors.

4.5.3. Quantitative Data Collection Process

Based on the information generated in the analysis of the qualitative data, an online questionnaire was constructed. The goal was to see if others who owned or controlled spatially-related data would agree with the responses of the ten interviewees regarding the hypothesis points. The author sent an invitation to participate in the research to listservs concerned with geographic information of different types, specifically to members of the Global Spatial Data Infrastructure Association and to members of the Maine Geolibrary listserv. In addition, printed flyers inviting participation were distributed at a conference of the Maine GIS User Group and the Maine Municipal Association.

Many users of spatially-referenced data are also creators of that data, as are many users and creators of other types of data or information on the World Wide Web. This phenomenon, dubbed “produsage” by Axel Bruns (Bruns, 2008). In this framework, those who both produce and use data are referred to as “producers.” This is similar to the situation in current media production tools where there is a line of products aimed at “prosumers,” people who both produce and also consume media products such as music or video, often in an online context.

Given this “produsage” tendency online, the survey instrument used the first question to separate those who were producers of data, or who had significant influence on data sharing in their organizations (potential contributors), from those who considered themselves only potential data users.

There was no attempt to ensure that data owned or controlled by respondents was locally generated or privately owned since the complications of trying to

pre-qualify potential respondents while simultaneously encouraging them to take a very short, “simple” survey was felt to be impractical. The fact respondents stated that they owned or controlled data rights and would consider making their data available in an online environment without financial compensation qualified them to respond to the survey.

The types of data that respondents owned or controlled included location and contents of waste disposal containers, location and types of health centers, vegetation distribution, land ownership, and many other types of data. While no residence location information was requested from respondents, a number mentioned their geographic locations, several of which were outside of the United States.

All of those who identified themselves as potential contributors also considered themselves potential users. If they completed the entire survey, they answered 20 questions. Of those questions, six requested text based answers. The other questions required either yes/ no responses, or responses rated on a 1 to 5 Likert scale.

Those who identified themselves as not owning or controlling data were asked to answer 11 questions, of which three requested text-based responses. They are not included in this study.

As in the qualitative portion of the research, the author made no attempt to construct a statistically valid sample of all potential contributors or users of an online commons repository since that universe is simply unknown. Rather, the goal was to gather a reasonable number of responses from self-identified potential contributors to either validate or invalidate the qualitative research

findings. Survey respondents were asked for no demographic or other potentially personally identifiable information, and were assured that all responses were anonymous and confidential.

There was a total of 197 click-throughs from the survey splash page to the actual survey instrument. Each click-through response was given a specific ID for analysis purposes.

Of 197 click-throughs, 120 identified themselves as owners/ controllers of data. Of those, 100 completed all questions, 10 answered some of the questions, 10 answered none of the questions. For all of the quantitative results discussed below, n=110 unless otherwise noted.

4.6. Results and Discussion

The interviews were recorded, transcribed, and then coded. Since all interviewees were asked the same set of questions, initial top-level codes were based upon those questions. Codes included conditions (which owners might put on use of contributed data); metadata (short description, key words, etc.); evaluation (valuable or not, amount of time that a contributor would spend, etc.).

As additional aspects of responses appeared, sub-categories for the major categories were added to make meanings more precise, and a few additional top-level codes added for topics that emerged.

Based upon the responses in the interviews, a set of questions were developed that could be posed in an online questionnaire to ascertain whether other potential contributors who completed all or some of the online questionnaire would support or not support the views of the interviewees. The

questionnaire responses were then tabulated and compared with the interview results.

We review the results by each hypothesis sub-part.

4.6.1. Hypothesis Sub-part (a):

*a simple, clear licensing mechanism would help motivate potential contributors to consider contributing their data to an online commons-type repository.*⁴

4.6.1.1. Qualitative Findings. Three interviewees said that licensing was not an issue for them or their organizations since they would not put any conditions on the use of their data if they were to post it online. However, two of the three added that while there was much data they would be willing to make publicly available with no restrictions, there was also some data they might not wish to share in a publicly available online environment. This was also true of several other interviewees as well. (See discussion on withholding some data below).

All of the other interviewees indicated that they or their organizations would want attribution if their data were publicly available online, although they recognized that it is difficult to control what people do with information once it is online. As one person noted: “yea, if they were to use it in a publication or on a web site, I would ideally like to see some credit for it but I am not going to worry about it too much because it is not something that I have a lot of control over.” None of the interviewees said they would absolutely withhold their data if

attribution could not be guaranteed but seven of ten indicated that attribution, along with a way to ensure that was given, at least in the first instance, would be desirable to them or their organizations.

Three respondents also indicated that while they would be happy to make their data available for non-commercial use. If users wanted to use the data in a commercial context, then they would want to be contacted and negotiate some type of compensation with a potential commercial user.

Half of the interviewees had a concern which no “some rights reserved” licensing scheme at present addresses, nor perhaps is it a concern that is addressable through licensing. They wanted some type of assurance that their data would be used properly. By “properly,” they meant slightly different things but the core concern was summed up nicely by one interviewee: “I think we would probably want to ensure some kind of conditions that protect the integrity of the data. I don't think we would be inclined to worry about commercial use or that sort of thing. I think we would be mostly concerned with are these data being used properly and are they not being taken out of context or are they potentially being used to misrepresent a situation where the data are not used in a way that we think are sensible or consistent.”

The same person indicated that if a user at home came across this group's data and misunderstood it, that would not be a serious cause for concern: “I think we would probably mostly be concerned about when and how the data is used in some kind of a publication. If someone is just sitting at their home computer and looking for data and drawing their own conclusions about things, I don't think we would be as concerned...I don't think we would attempt to try

to control every pair of eyes looking at that data, saying oh no you are not understanding this properly. I think the concern would be a newspaper article...”

The issue for those with what we might call a “downstream quality control concern” is that once their data is out of their control, it might be “corrupted or somehow altered and misrepresented,” as another interviewee put it. Even those who were not concerned about attribution and did not see any reason to put a license on the use of their data shared a concern that the data could be misused. One interviewee spoke of putting an “advisory” on the owner’s data that said, in this particular case: “don't use irresponsibly. That is, don't go to these particular zones and stress the birds.”

In some cases, the concern was so strong that it resulted in interviewees reporting they would choose to withhold data out of fear that it would be used improperly and/or misinterpreted, or was so sensitive that releasing it without knowing who might use it could have adverse effects. These were mainly cases of land trusts or other environmental organizations. In some cases, they had developed information about locations of endangered species. In other cases, they had negotiated easements or other land use agreements with landowners which the interviewees felt could create problems either for the land owners or for the organizations if they were made available to the public.

None of the questions in the original interview protocol spoke specifically to this concern. It emerged in three interviews during a general discussion of what conditions, if any, potential contributors might place on the use of their data in an

online commons environment. As a result, we added a specific question about types of data potential contributors might choose to withhold to the online questionnaire.

While this concern arose spontaneously among this particular group of interviewees, it has been a concern in institutional settings, for example, among cultural institutions (Eschenfelder and Caswell, 2010). That concern is becoming more acute in the online world.

4.6.1.2. Quantitative Results: Results from response to the online questionnaire are largely consistent on this topic with those gleaned from the personal interviews.

Respondents were asked to reply to a series of questions that began with:

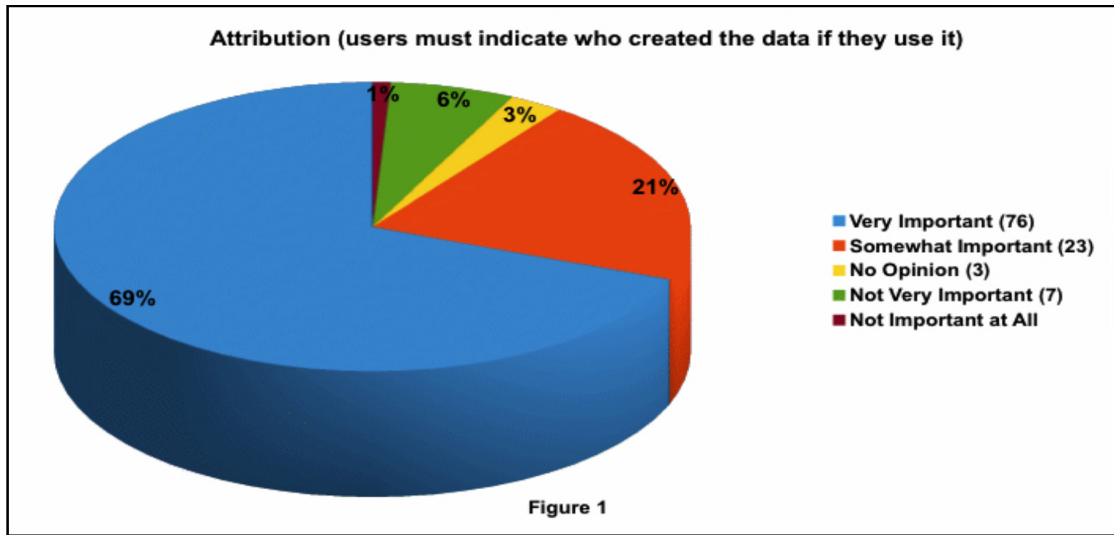
“If you were to consider making your data available online so that others could access and/or use it, please indicate how important each of the following would be in your decision whether or not to make your (or your organization's) data available.”

110 respondents indicated that they had data that they might consider making available online and answered at least one other survey question.

Respondents were asked to rate each item on a scale of 5–“Very Important,” to 1–“Not Important At All.”

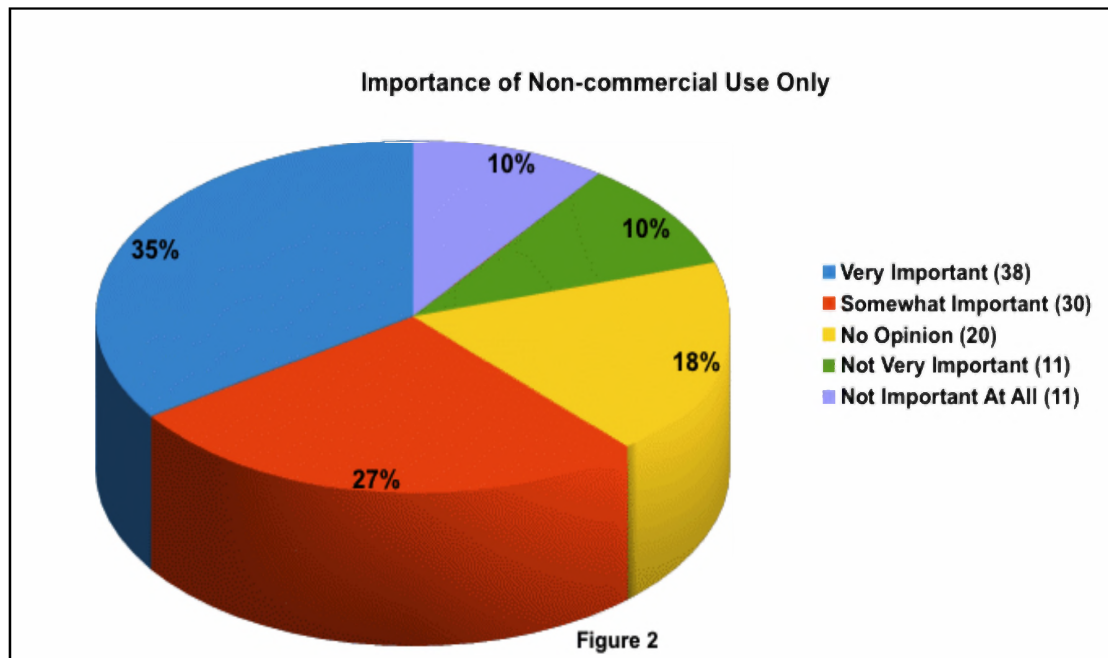
The first item concerned “Attribution.” Note that these and all following percentages are rounded. The raw number is noted next to each response description, the percentage indicated in the graph.

Figure 1: Attribution.



The question of non-commercial versus commercial use of contributed data arose in the interviews. As a result, a specific question addressing that issue was included in the questionnaire. Respondents were asked how important being able to make their data available for non-commercial use only would be to them.

Figure 2: Importance of Non-Commercial Only Use

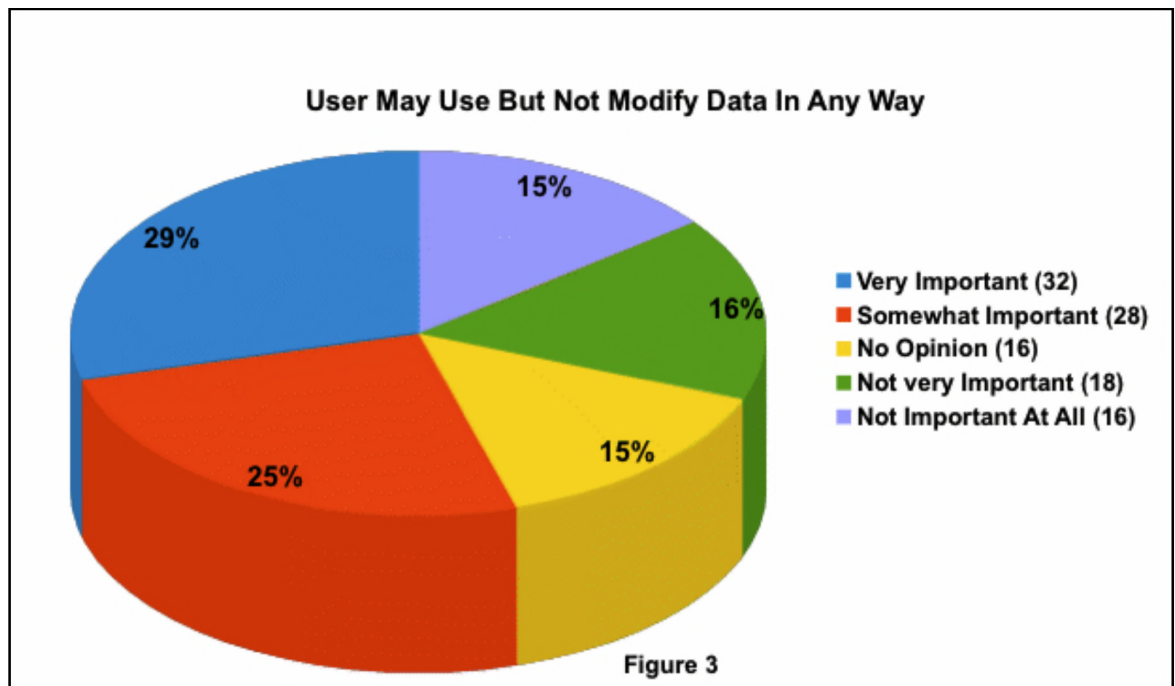


Only three of ten interviewees specifically mentioned non-commercial use as a use concern. This differs from the 62% of respondents to the questionnaire who would find being able to specify non-commercial use to their data Very or Somewhat Important. This discrepancy could be due to the fact that there was no specific question about non-commercial use asked of the interviewees, only general questions regarding any conditions they might put on the use of their data. The fact that three interviewees spontaneously mentioned this concern led to it specifically being included in the questionnaire. It is interesting to note that the 35% of questionnaire respondents who considered it “Very Important” to be able to indicate use of their data was only for non-commercial purposes matches reasonably well with the 30% of interviewees who spontaneously expressed this concern.

Two other concerns arose during the qualitative analysis of the interview data and both were included specifically in the questionnaire.

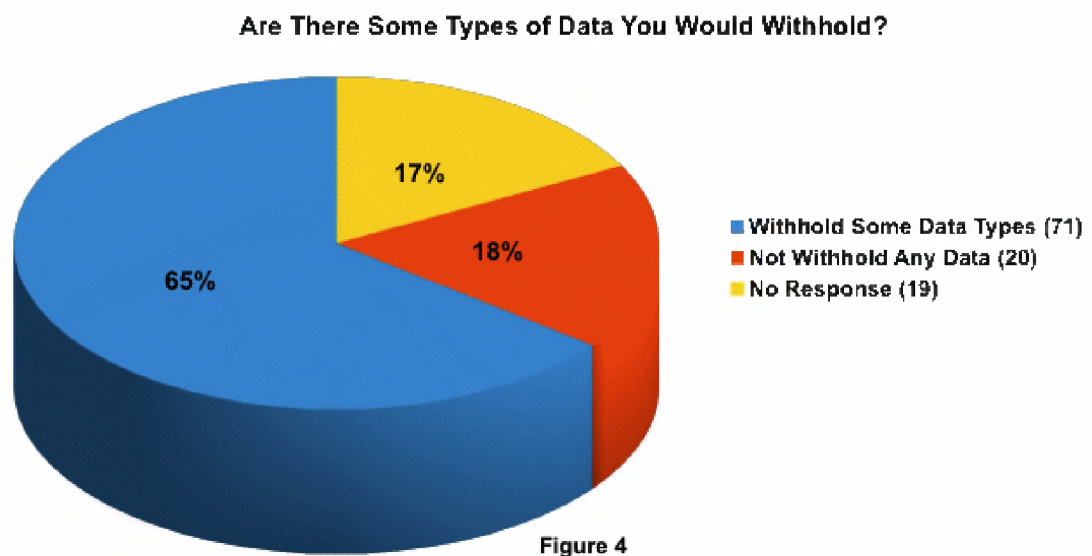
The first involved concerns about data being corrupted or misused because of a lack of understanding. As noted above, there is no license of any sort that can guarantee that data will not be misunderstood. However, there are “some rights reserved” licenses which prohibit modifying the data as a condition of the license grant. Therefore, respondents were asked how important being able to specify “User may use the data but not modify it in any way” would be.

Figure 3: User Modification of Data



Over half of respondents seemed to share a concern that their data not be manipulated. We did not ask specifically why, but it is likely that concerns over data integrity and possible corruption of data, as revealed in the interviews, may also have been a concern of the questionnaire respondents.

Figure 4: Types of Data That Might Be Withheld



As noted, the second concern that arose during the interview phase of the research involved withholding some data which potential contributors considered sensitive. To explore this issue further, the following question was included in the online survey: "Is there any type of data which you possess that you would NOT be willing to make available in an online commons-type repository? If so, please briefly describe it and indicate why you would not make it available."

Since the questionnaire group was much larger than the interviews group, the types of data and rationales for withholding some data varied across a broader range than those mentioned specifically during the interviews. The bulk of the reasons for holding data back mentioned by questionnaire respondents fell into the following categories:

- Homeland Security;
- financial privacy, e.g., tax, income, property information;
- personal privacy, e.g., health related information;
- some part of data purchased from or held by another owner;
- endangered or sensitive species information;
- incomplete data or not of high quality;
- high level of expertise required to understand properly and thus could be misinterpreted;
- part of ongoing academic research and researchers do not want to be “scooped” on their research; and
- hope of generating future income or cost reimbursement.
- Among these reasons are all of those expressed by interviewees, as well as a number of additional ones.

4.6.2. Hypothesis Sub-part (b):

a simple process for attaching descriptions to the data. The goal would be to make the data easier for users to discover.

4.5.2.1. Qualitative Findings: Metadata is often the weakest part of data management. Developing full metadata descriptions using the Content Standard for Digital Geospatial Metadata (CSDGM), the Federal Geographic Data Committee standard, involves dealing with over 300 fields, as does the international ISO-19115 *Geographic Information-Metadata* standard. Few professionals, and almost no non-professionals, even attempt to provide complete metadata descriptions for spatially-referenced data sets. Yet using metadata that conforms to international standards is key to making data widely visible in an organized way. This is in contrast to, for example, non-standard tagging in applications like Google Earth or Flickr which international search protocols such as OAI compliant search tools are not able to harvest and make available.

Potential contributors to an online commons environment would not be expected to create standards based metadata for 300 fields. However, there is a more limited set of ISO-19115 core metadata items which would be practical to have contributors provide, and which could be done in a few minutes without the contributors having any knowledge of metadata or of metadata standards.

During the interviews, we were interested in discovering whether interviewees had already developed basic metadata, i.e., descriptions of the data file contents and keywords that could serve as finding aids in an online

environment. If they had not, we inquired whether they felt it was worth investing time and resources to do so, and how much time they would be willing to invest to provide such information.

None of the ten interviewees had provided either short descriptions of the files that contained their data nor had they attached any keywords to the files. All of the interviewees were aware of the usefulness of metadata but none had found a compelling reason to create either file descriptions or keywords for their files. As one interviewee jokingly put it: "I am an evil person, I have not done the metadata."

This absence of metadata did not cause any operational difficulties locally since the data was either owned and used by an individual or by a very small group of people who all knew what the data was about or could simply ask a colleague if they did not. None of the data was online at the time of the interviews so making it more discoverable had not been a priority.

All of the interviewees recognized the value of having useful metadata in an online environment, and all could quickly identify keyword terms that would be appropriate for their data.

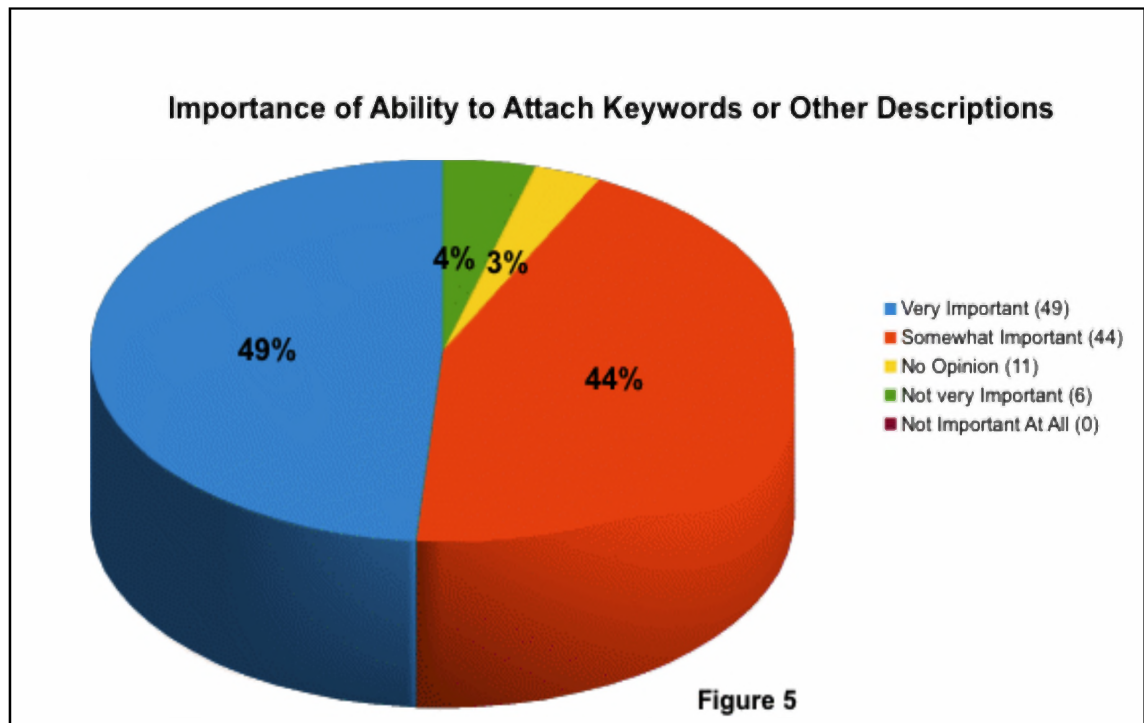
The question of how much time they might be willing to invest in creating metadata for their data files if the files were to be placed in an online environment varied. In most cases, interviewees felt that since they were individuals or worked with very small organizations, they would have to believe that there would be a use for their data. They then would have to evaluate for themselves or with their boards or colleagues, in the case of organizations, whether investing that time would further their missions or purposes.

Even with that caveat, eight of the ten interviewees would be willing to dedicate from a half-hour per file to “as long as it takes” to provide file descriptions, keywords, and location information for their data. The other two respondents felt that once they had set up a system, the nature of their data was such that it would take only five minutes or so to provide that information per file.

In sum, all interviewees recognized the value of providing metadata for their files if their data were to be placed in an online environment, and they would be willing to dedicate time and resources to do so if they were convinced that others might value and use their data, and that the knowledge required to input the information was minimal. However, none of the interviewees had actually already created metadata in the offline environments in which they worked at the time of the interviews.

4.6.2.2. Quantitative Results. Questionnaire respondents were asked how important the “Ability to attach keywords or other descriptions to your data so that further users could find it more easily” would be in an online commons-type environment.

Figure 5: Importance of Ability to Attach Descriptions.



Of the 110 respondents that answered the previous question, 102 also answered a text question asking them how much time they would be willing to devote to uploading and describing their data. As with the interviewees, the spectrum was wide, ranging from five minutes to “as much time as it would be necessary to do so.” A few respondents said they had already created metadata and one said the process would be automated. The great majority of those responding indicated they felt that metadata for their data was important and that they would devote the time necessary to provide it.

As with the interviewees, questionnaire respondents strongly recognized the value of adding metadata to their files if they were to make them available online, and almost all would be willing to take some time to provide metadata.

4.6.3. Hypothesis Sub-part (c):

a simple post-publication peer evaluation mechanism that would both provide feedback for contributors, and provide information on quality and suitability for use for users.

4.6.3.1. Qualitative Findings. Nine out of ten of the interviewees viewed the ability of users to comment on data to be a positive factor in potentially placing their data in an online commons type environment. The other interviewee said that it would not make much difference because “I don’t necessarily know why but I would tend not to trust, you know, people’s review of my data.”

The others, however, saw that capability as a definite plus. There were suggestions that there be some sort of registration system so that commenters would be registered, even if they used a screen name rather than their own name, to minimize abuses of an open commenting system. Interviewees also indicated that, if possible, they would like to know something about the commenter’s use of their data to help them judge whether the comment was appropriate to their data. Interviewees felt that they had developed their data for a particular type of use and they were interested in receiving feedback on it when it was used in a similar context. Several also indicated an interest in being able to contact a commenter if what the commenter said could be helpful for improving their data or suggested an additional use.

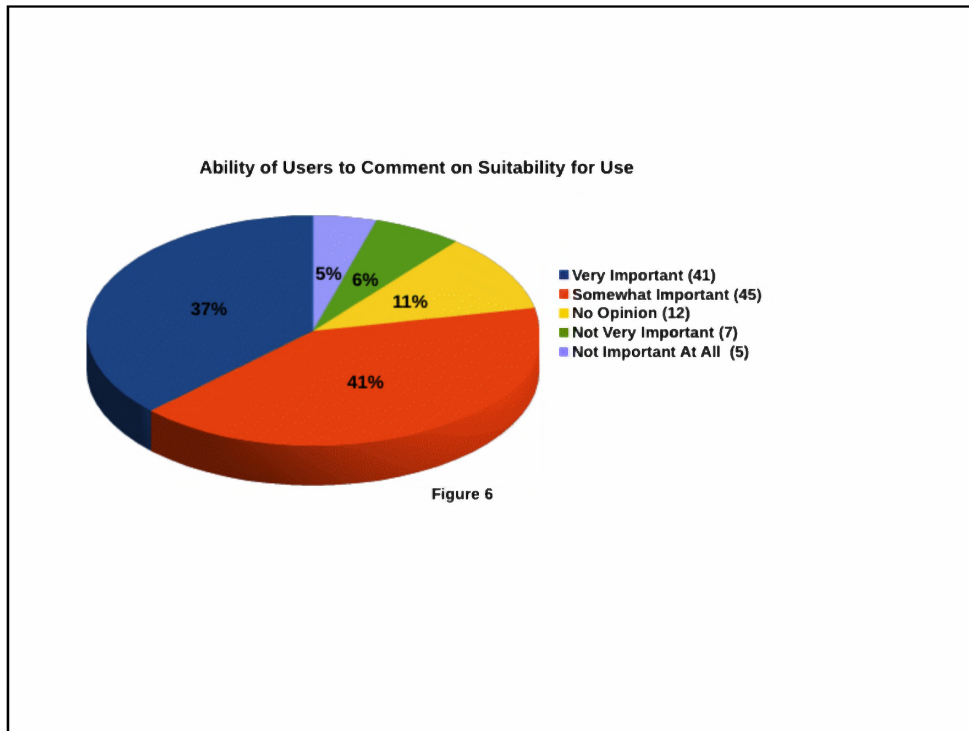
The advantages of user feedback from a data contributor's perspective included knowing that someone else had found their data useful for particular purposes, receiving suggestions or questions that they might not have thought of themselves, and using comments by users to improve their data.

One additional positive mentioned by two of the interviewees highlighted the value of knowing that one is part of a larger community with similar interests: "...the connectivity, the sense of networking and the sense of camaraderie almost that sharing information could provide or does, at least on paper, seem to provide is in itself a good, it is a social community kind of good and that to get some feedback that says, 'hey, we are using your data' would feed that sense that you're part of something bigger than your own effort. And I think that would be helpful and inspiring so to be able to get that feedback, you know, you have to have some venue where that can happen." In this person's opinion, a peer commenting mechanism could support that sense of community, especially for those working in small non-profit organizations.

Interviewees found a peer evaluation/ commenting system to be a very desirable characteristic for an online commons-type data environment.

4.6.3.2. Quantitative Results. Questionnaire respondents who identified themselves as owning or controlling data overwhelmingly felt that the "Ability of users to comment on the suitability of the data for their uses" would be important.

Figure 6: Ability of Users to Comment on Suitability for Use



Although the questionnaire did not ask for reasons why this capability might be important, the numbers support the overall consensus of the interviewees that a commenting/evaluation capability would be valuable from the perspective of potential data contributors.

4.6.4. Repository Maintenance

While not a specific sub-part of the hypothesis, interviewees were asked in a general way about desirable repository characteristics: “Would it make sense to you to make your data available in a central location on the web so that people who might wish to use your data could do so without contacting you directly?”

If the answer was yes, the follow-up question was “could you describe any characteristics of such a central location that would encourage you to make your data available there?”

4.6.4.1. Qualitative Findings. In response to this question or in other parts of the interviews, several interviewees brought up concerns about the nature of a hypothetical online commons repository. While they recognized potential value in such a repository, they also realized that it would take effort by themselves or their organizations to prepare and upload their data. As one person noted: “it would take a huge effort for us to get it into a consistent format to upload it...” While interviewees were open to making that effort, they felt that there should be certain assurances about the repository to justify the work involved.

One concern focused on how such a repository might look to users and whether there would need to be different sections, e.g., a section specifically for student generated data so users would know the data might not be of professional quality. There were also comments about whether or what kind of guidelines for responsible use of the data there might be. But the largest operational concern was the longevity of such a repository.

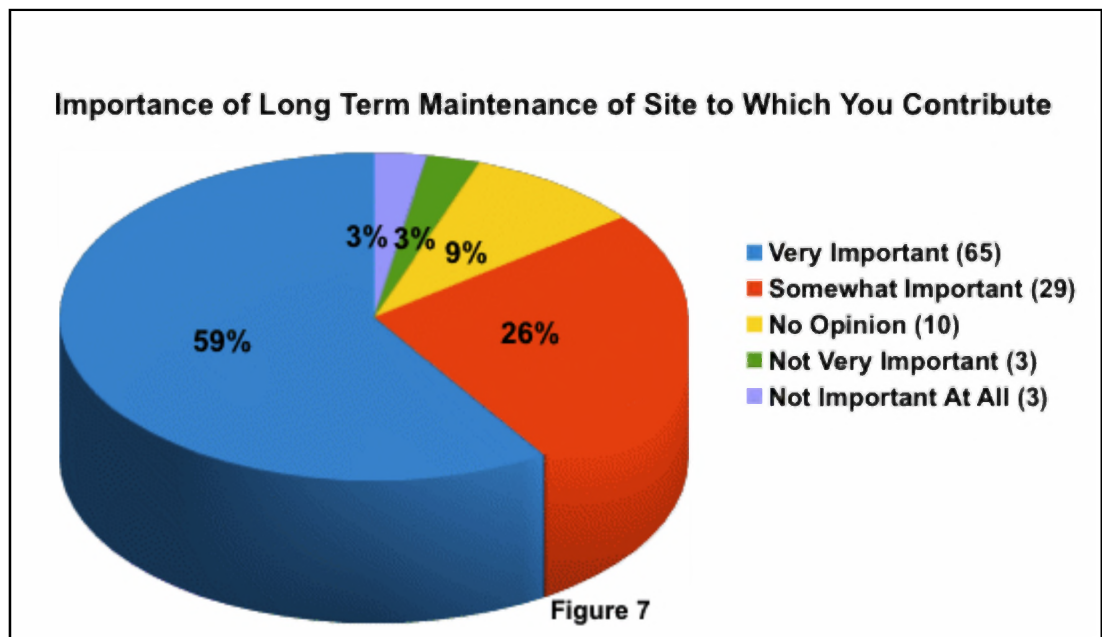
Since almost all of the interviewees indicated it would take additional work to prepare and upload their data, most felt that there would need to be some assurance that the repository would be maintained over time if they were to make the effort necessary to contribute their data. One interviewee expressed the concern in these words: “I fairly frequently see this ‘start up, some interest, and then you know, decline’ profile and because of that I guess I tend to be a little

nervous about starting up or being part of the start up because I don't know whether my efforts at the front end are going to result in the kind of long-term engagement that I was anticipating or hoping for."

Based upon the strength of this concern about the longevity of a repository, a question was added to the survey on this topic.

4.6.4.2. Quantitative Results. Survey respondents were asked how important "Long term maintenance of your data on the online site" would be to them. Overwhelmingly, long term stability matters to potential data contributors.

Figure 7: Importance of Long Term Repository Maintenance



This response mirrors the response of the interviewees as to the importance of long term maintenance of any commons type online environment.

4.7. Chapter Conclusions

Based on the interview conversations and analysis and the online survey results, the hypothesis put forth above seems to hold. Results from the interviews are generally confirmed by the survey results. Although in some cases percentages differ, concerns were consistent overall both in the interviews and the survey responses.

The purpose of this research is to provide guidance to those who may wish to construct a commons-type repository in which anyone could make their data available for sharing with others, although the results could be of use in institutional repository and other settings as well.

This research, subject to the caveats listed below, suggests that it would be desirable from the perspective of potential contributors of data to provide infrastructure capability that would:

- allow users to attach conditions to the use of their data,
- provide basic information that could be translated into standards based metadata, and
- receive comments and feedback from users.

Assuring potential contributors that such a repository would have staying power and that their data would be available over time would also be an important consideration for potential data donors.

4.7.1. Limitations

This research has several limitations. It does not purport to be a statistically valid sample of potential contributors. That universe is simply not known nor

probably knowable. Respondents to the online survey were self-selected. While interviewees all had spatially-related data that was generated locally and not available online at the time of the interviews, no such claim can be made for the survey respondents, although respondents were invited to participate only if they might be willing to make their data available without up-front financial remuneration, and only if they owned or controlled spatially-referenced data.

These limitations prevent any assertion that the hypothesis is “proven” but they do not, we feel, limit the usefulness of the research results for their intended purpose: to provide guidance to those who may in the future choose to construct an online commons for spatially-referenced data that anyone, non-professional and professional alike, can contribute to with no special expertise. Such a commons could help to make visible much currently invisible data for the benefit of all.

In that regard, the author hopes this research has something to offer.

Chapter Notes

1. See, for example, the Atlas of Canada (<http://atlas.gc.ca/site/index.html>), and Geoscience Australia (www.ga.gov.au). In the U.S., initiatives such as the National Map (<http://nationalmap.gov>), the National Atlas (www.nationalatlas.gov), and Geo.Data.Gov (<http://geo.data.gov/geoportal/catalog/main/home.page>) serve similar functions. They generally contain a wider array of data since in the U.S., the federal government cannot hold copyright on materials it generates. Similarly, there are non-governmental disciplinary and special purpose repositories that

exist to capture large scale spatially-referenced data, e.g., PANGAEA (<http://www.pangaea.de>) and OneGeology (<http://www.onegeology.org/>). An example of a global interface for accessing earth observation data sets and services is the Global Earth Observation System of Systems (GEOSS) (<http://www.earthobservations.org/geoss.shtml>).

2. Under United States copyright law, facts themselves cannot be copyrighted but original arrangements of facts can be. For the purposes of this research, we assumed that data sets owned or controlled by interviewees and questionnaire respondents included sufficient original arrangement to qualify for copyright protection although this is undoubtedly not true in all cases. A simple list of dates and temperature readings at a particular location on those dates, for example, would probably not qualify for copyright protection. Rather than muddy the water by trying to make determinations of copyright status of particular data sets, we assume all potential contributions would qualify for copyright protection.

3. Ten interviewees in a qualitative study is a large enough number in qualitative studies to get a good sense of qualitative attitudes of, in this case, potential contributors to an online commons environment. The same is true in other interactive intensive studies such as software usability studies (Hwang and Salvendy 2010).

4. To get a sense of one possible approach, see Campbell et al. 2006.

CHAPTER 5

DESIRABLE CHARACTERISTICS OF AN ONLINE DATA COMMONS FOR SPATIALLY REFERENCED, LOCALLY GENERATED DATA FROM DISPARATE CONTRIBUTORS

5.1. Background

A significant body of spatially-referenced, locally-produced data developed for specific local purposes exists on the hard drives and backup systems of individuals, nonprofit groups, private associations, universities, private companies, and other nongovernmental organizations across the United States. Spatially-referenced data, as the term is used here, is data that refers to a particular physical location. Examples might include a university botany class project that locates and catalogs all the trees more than 15 feet tall in a small town; a homeowners' association that monitors the water quality and plant growth of the lake on which members' properties are located; a land trust that records environmental easements; or a historical museum that ties its photographic images to their physical locations, among many others.

In all these cases, the data gathered by these small local originators could be of great value to others if its existence were known. At present, however, very little of this data is available from a practical perspective to other scientific researchers and potential users. It is, for all intents and purposes, completely or partially "invisible."

While much emphasis has shifted in recent years to providing geospatial services, there still is a strong need for service developers to be able to find and exploit existing geographic data that would make those services more effective and efficient. Many efforts at the national and state levels are being made to make government-generated spatially-referenced data available to the public. In the United States and in other countries around the world, initiatives are under way to make geographic information more freely available to scientists and to the general public. In English-speaking countries, for example, UK Location (<http://location.defra.gov.uk>) in the United Kingdom, the Atlas of Canada (<http://atlas.gc.ca/site/english/index.html>), and Geoscience Australia (www.ga.gov.au) provide open access to some government-generated spatially referenced data. In the United States, initiatives such as the National Map (<http://nationalmap.gov>), the National Atlas (www.nationalatlas.gov), and the geospatial section of data.gov (<http://www.data.gov/geospatial/>) serve similar functions. These U.S. sites contain a wider array of data than many other national portals because the U.S. federal government cannot hold copyright on materials it generates, and because some state governments make their state-level data visible through these gateways. Efforts also are under way to make international sharing of large datasets more viable, especially with regard to divergent approaches to data licensing and use rights (Onsrud et al. 2010). GEOSS Data Collection of Open Resources for Everyone (GEOSS Data-CORE 2014) is an example of an international initiative to support open access to geographic data gathered by governments across nine societal benefit areas (GEOSS 2014).

Similarly, disciplinary and special purpose repositories exist to capture large sets of spatially referenced data. Examples include PANGAEA (<http://www.pangaea.de>), and OneGeology (<http://www.onegeology.org>).

Google Maps, Google Earth, Virtual Earth, and Open Street Maps provide structured environments where the user may take advantage of a data-gathering and display infrastructure to contribute data or volunteer effort to a commercial or open-data environment. In these information infrastructure environments, legal and data management issues as well as data format issues are closely controlled by the infrastructure system provider. These are not infrastructure environments for depositing or finding diverse geographic datasets, and this article does not address such environments.

We conclude that no gateway exists analogous to the Global Earth Observation System of Systems (GEOSS) that could provide more visible and efficient access to millions of spatially referenced datasets drawn from disparate locally generated sources. Note that the GEOSS is a *portal* or *gateway* for finding relevant geographic data and services rather than a *repository* of geographic data itself. Furthermore, the metadata on geographic data and services contained within the GEOSS is provided or mined from primarily national and international government members and participating organizations of the Group on Earth Observations (GEO). The GEOSS serves as an exemplar of the kind of infrastructure that can make geospatial data files and services from widely disparate cooperating sources much more readily findable.

5.2. Volunteered Geographic Information

In the past decade, regular people have become producers as well as consumers of geospatial data, a phenomenon variously called neogeography (Turner 2006, Sui 2008), ubiquitous cartography (Gartner et al. 2007), collaboratively contributed geographic information (Bishr and Mantelas 2008), and volunteered geographic information, or VGI (Goodchild 2007). VGI seems to be the most widely used term at present.

Affordable, portable GPS devices have made it possible for anyone to make a quite accurate observation of the position of an object on the face of the earth. Simple-to-use infrastructures that use Google Maps, Open Street Maps, or similar frameworks make it easy to add those observations to a map, and to attach notes or information to the location. To date, the great bulk of VGI activity has involved this form of adding locations and labels of features within a mapping facilitation framework or to already existing maps. At the observation level, then, VGI contributors can contribute data in many situations as well as trained geographers could in pre-GPS days.

Adding or correcting locations, names, and characteristics of features on a map base such as Google Maps or Open Street Maps is a type of spatially referenced data but there are many other types including complete datasets of various kinds such as the examples mentioned previously. Most of the examples involve “asserted” rather than “authoritative” data (Bishr and Mantelas 2008). In VGI-contributed environments, where disparate datasets are only asserted as potentially useful and not vouched for, context becomes crucial. VGI data, or any data, collected for one specific purpose may not be relevant or useful or even

accurate for a different purpose. Potential online environments that may feature collections of data generated locally for disparate purposes need to contextualize that data for the data to be useful.

5.3. Desirable Characteristics of an Online Spatially Referenced Data

Repository

Simply having an online gateway or home for widely disparate, spatially-referenced, locally-generated datasets could be of significant use for providing access to this type of data. It probably would be of greatest use to geospatial specialists and professionals desiring to find and draw from existing spatially referenced data to provide further products and services. We refer to this perceived online gateway or home as a Commons of Geographic Data (CGD). However, if such a facility or capability, centrally located or distributed, is to be of maximal use over time to both professional scientists and to interested nonprofessionals, a number of studies and reports suggest that it should include functionality that enables users to know usage rights and search for and discover data using standards-based metadata, and provide users with a way to access evaluation commentary from previous users of the datasets and offer comments of their own. See these common elements in, for example, Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content (IMLS 2003), Licensing Geographic Data and Services (NRC 2004), and To Stand the Test of Time: Long Term Stewardship of Digital Data Sets in Science and Engineering (ACRL 2006).

In a commons-type environment for data users, data is made available under a license—if a license is necessary to use the data—that grants permission for use as long as any stipulated conditions are adhered to. This makes it possible for potential users to be sure that they may use any data found in such a commons environment without seeking additional permission from the owner. In such environments, permission already has been granted as long as any conditions specified in the license are respected. Creative Commons licenses are one example of so-called “some rights reserved” license types typically found in a commons environment for materials that are not in the public domain. Creative Commons licenses currently are used in more than half a billion digital works. Creative Commons and its affiliate, Science Commons, have designed several licenses specifically applicable to datasets (Creative Commons 2014) that could be used in a Commons of Geographic Data.

An online Commons of Geographic Data with the characteristics listed previously does not exist at present. If such an environment were contemplated as a future project, based on the reports previously cited, important questions arise almost immediately. If there were such an online data commons repository for small, privately generated datasets, would people who are interested in spatially referenced data be willing to access and use the data in such a repository? What type of functional characteristics of such a repository or gateway would help to motivate those potential data users to actually examine and possibly use the data located there for their own purposes?

It may seem reasonable to assume that such characteristics would be desirable to potential users, but at this point in time, reasonable or not, this still is an assumption. The goal of this research is to address this question empirically.

5.4. Hypothesis

The purpose of this research is quite practical. It is hoped that the results may provide some guidance for future architects of an online Commons of Geographic Data about functionality that potential users would be interested in finding in an online commons environment for spatially referenced small datasets from disparate sources, if and when such a commons environment is constructed. The results could suggest several areas for future research, and might also be of use to those who currently operate data gateways or repositories that they would like to make more responsive to users' interests.

Based on common elements in the reports noted previously as well as in other data-preservation related studies (e.g., Committee on Science, Engineering, and Public Policy (U.S.) 2009, Interagency Working Group on Digital Data 2009), we hypothesized that potential data users would be willing to consider using data accessed through an online gateway or data repository if such a facility included:

- (a) a simple, clear licensing mechanism that reveals ownership of, and conditions for use of, the contributed data;
- (b) a simple, effective searching/ finding mechanism that provides an option to search using either *Thesaurus*-controlled vocabulary, "plain English" keywords, or location; and

(c) a simple postpublication peer-evaluation mechanism that will provide information on quality and suitability for purpose for users.

5.5. Method

To test this hypothesis, we used a combination of qualitative and quantitative research procedures (Onwuegbuzie and Leech 2004; Ragin, Nagel, and White 2004). Personal interviews were conducted with ten people who were regular users of spatially-referenced data. These particular interviewees also were generators of spatially referenced data. The findings from these qualitative interviews were used to construct an online questionnaire, and results from that questionnaire with responses from a much larger group (139 people) were compared with the results from the interviews to see if the qualitative results were supported by quantitative data.

5.5.1. Methodological Limitations

The respondents in this study are not in any way meant to be considered a statistical or otherwise representative sample of potential data users of an online commons gateway or repository for spatially referenced datasets from disparate sources. The major reason for not attempting to select a representative sample of potential users is that the universe of such users is unknown and probably unknowable. Thus, the combination of qualitative in-depth interviews with quantitative data was chosen to produce findings that would be informative, even though not “proven” in a statistical sense, for future designers of an online

commons-type geospatial data environment, and that could suggest directions for future study.

All participants in the study were self-selected. In addition, to generate quantitative responses online, given the reverse traceability of personal user information in today's online environment, potential respondents were guaranteed anonymity by requesting no geographic, employment, or other demographic information. This makes some types of statistical analysis impossible.

5.5.2. Interviewees and Data Types

Interviewees were selected based on a “snowball technique” (Maxwell 2005). Interviewees were referred by word of mouth from those interested in spatially referenced data who were located in geographic areas accessible to the authors. Those who agreed to participate were asked if they could recommend others who might be potential interviewees. In the final group of ten interviewees, seven were from Maine, one from Massachusetts, one from Pennsylvania, and one from North Carolina.

One interviewee was a graduate student working on a spatial-data research project; one regularly dealt with spatially referenced data as part of the respondent's employment, although the role the respondent held in this study was as a volunteer citizen on a municipal committee. About half the respondents were familiar with and used GIS software to a greater or lesser degree; about half did not. Four were involved with land trusts of one type or another, one was an author of nature books, one a high school teacher, one a local museum curator,

and the others were involved with other types of local civic groups. All the spatially-referenced data that these originators were gathering were deemed by the investigators and the gatherers to be of potential interest to others in the future but none of the data was available on the Web.

5.5.3. Qualitative Data-collection Process

The purpose of these qualitative interviews was to test whether the hypothesis above would hold, and to discover if other important desirable characteristics arose spontaneously in the interviews. All interviews were conducted from the same interview instrument by the same interviewer. The interviews were transcribed and coded, and then the transcripts were checked against the voice recordings for accuracy. A summary of key points then was sent to each interviewee for correction, if necessary, and for confirmation. None of the interviewees who responded submitted any corrections other than spelling errors.

Because all interviewees were asked the same set of questions, initial top-level codes were based on those questions, e.g., “conditions” (which owners might put on use of contributed data); “metadata” (short description, keywords, search order, etc.); “evaluation” (valuable or not, amount of time willing to spend commenting, etc.). As additional aspects of responses appeared, subcategories for the major categories were added to make meanings more precise, and a few additional top-level codes added for topics that emerged that were not specific responses to asked questions but that were relevant to overall online data commons use.

5.5.4. Quantitative Data-collection Process

Based on the information generated in the analysis of the qualitative data, an online questionnaire was constructed to see if others who identified themselves as users of spatially referenced data would agree with the responses of the ten interviewees regarding the hypothesis points. Notice of the existence of the questionnaire along with an invitation to participate in the research was sent out to listservs of those concerned with geographic information of different types, specifically to members of the Global Spatial Data Infrastructure Association and to members of the Maine Geolibrary listserv. In addition, printed flyers inviting participation were distributed at a conference of the Maine GIS User Group and the Maine Municipal Association.

The survey instrument used the first question to separate those who were owners of, or who had significant influence on data sharing in their organizations (potential contributors), from those who considered themselves only potential data users.

All those who identified themselves as potential contributors also considered themselves potential users, and there were additional respondents who considered themselves users only. We report on the results of the questions answered by all users, including those who also identified themselves as owners or controllers of spatially referenced data. There were 11 questions data users were asked to answer in the survey, of which three requested text-based responses.

As in the qualitative portion of the research, no attempt was made to construct a statistically valid sample. Rather, the goal was to gather a reasonable number of responses from self-identified potential users of spatially referenced data to either support or invalidate the qualitative research findings.

There was a total of 197 click-throughs from the survey splash page to the actual survey instrument. Each click-through response was given a specific ID for analysis purposes. Of 197 click-throughs, 139 completed some or all of the questions put to users.

5.6. Results

We review the results by each hypothesis subpart. Although the prior discussion refers to both portals and repositories for geographic data, with the human subjects we focused on the simpler concept of data repositories. However, we believe the results are generalizable for also guiding feature developments for portals or gateways such as GEOSS that lead to distributed repositories or portals.

5.6.1. Hypothesis Subpart (a): Simple Clear Terms of Use

Data users would be willing to consider using data in an online data repository if such a repository included a simple, clear licensing mechanism that reveals ownership of, and conditions for use of, the contributed data.

5.6.1.1. Qualitative Findings. All ten of the interviewees indicated that they would want to be able to check license conditions before they decided to download and

use data, and that they would respect any conditions that were put on the use of the data in a particular file. Most indicated that they would want a simple-to-understand statement of what they could or could not do with a data file. In the words of one interviewee: “I would want to be able to identify the conditions or at least get a sense of the conditions very quickly . . . I am not going to spend a lot of time reading a three-page license agreement.”

Several assumed that any conditions for use would be stipulated when a file was found, and certainly by the time it was opened, although another interviewee said that the interviewee always scans the Web page a file appears on to see if, for example, attribution is required.

Several interviewees referred to ethical considerations when describing whether and why they would check any licensing conditions before using the data in any but a personal way. Two of the interviewees indicated specifically that they would not bother to check for licensing conditions if they were just looking at the data for their own information, but if they contemplated using it in any additional way, they would check and respect any conditions of use.

Interviewees were asked if the presence of conditions of use that were clearly stated before opening a file might impact whether they would choose to look at a data file or not. Responses were evenly divided between those who would look at the data anyway and those who would not bother if they felt the conditions would preclude the use that they might wish to put the data to.

5.6.1.2. Quantitative Results. Results from responses to the online questionnaire are consistent on this topic with those gleaned from the personal interviews.

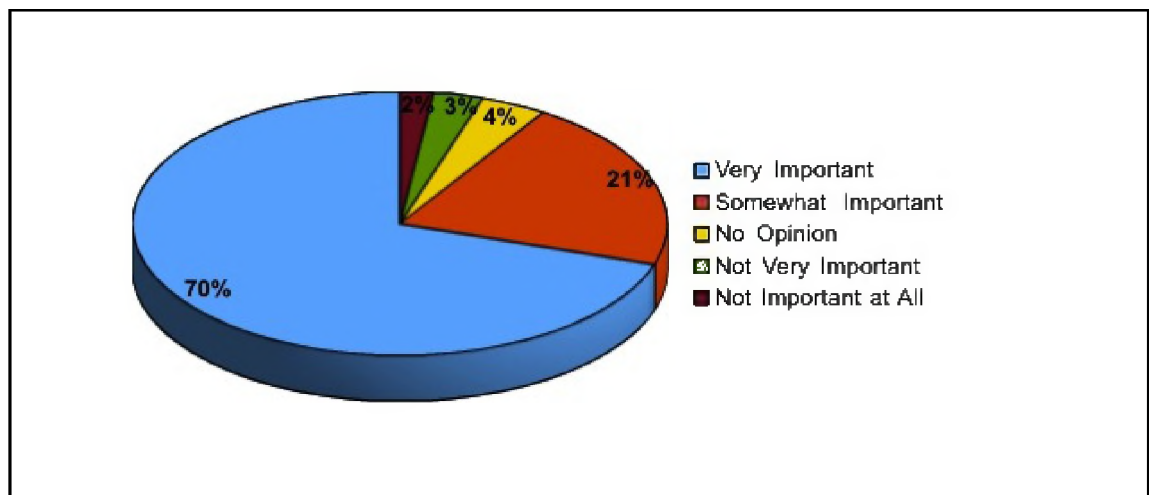
Users were asked in each question “If you were looking for data that others had contributed to an online commons-type environment, please indicate how important each of the following would be in your decision of whether to access and/or use such data . . .”

Users were given five choices:

- Very Important
- Somewhat Important
- No Opinion
- Not Very Important
- Not Important at All

This first question asked how important it would be that “Conditions for the use of the data are clear.” (Note that all the following chart percentages are rounded.)

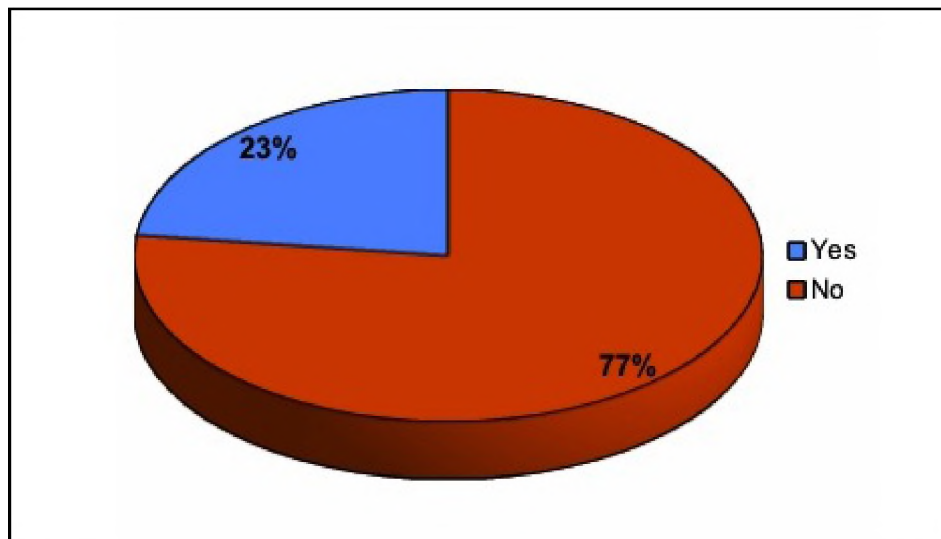
Figure 8. Importance of knowing conditions for use for data (n=139)



The importance of knowing the conditions for use expressed by interviewees is mirrored in the larger population of questionnaire respondents, with 91 percent indicating that such knowledge would be “Very Important” or “Somewhat Important” to them.

Addressing the question of whether licensing conditions put on the use of the data would affect potential users from accessing the data, respondents were asked: “If conditions for use of the data were clear, e.g., requiring attribution or noncommercial use only, might there be any conditions that would prevent you from examining the data?”

Figure 9. Would any conditions prevent you from examining data? (n=139)



Of those questionnaire respondents who responded “Yes” to this question, examples of conditions that might prevent users from examining a data file varied. The predominant response concerned limitations on commercial use.

Some other reasons included cost, administrative requirements, concern about data quality, limited bandwidth that would preclude downloading large files, and inability to modify the data for their own use.

5.6.2. Hypothesis Subpart (b): Search Mechanism

Data users would be willing to consider using data in an online data repository if such a repository included a simple, effective searching/ finding mechanism that provides an option to search using either *Thesaurus*-controlled vocabulary, “plain English” keywords, or location.

5.6.2.1. Qualitative Findings. None of the interviewees said that they would search for data based on *Thesaurus*-controlled vocabularies. All would begin searches using either natural language keywords and phrases, or location terms. All interviewees indicated that they might use either strategy first depending on what they were looking for at a particular time. About half indicated that they usually would begin with topic keywords, about half with location. However, each group then would use the other strategy to help narrow their results.

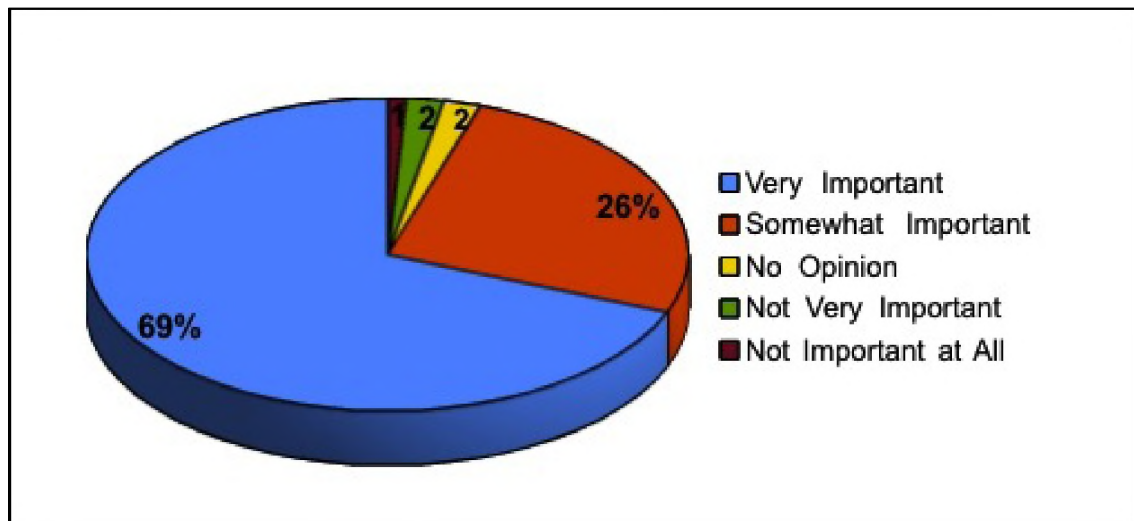
For example, an interviewee who served on a municipal recreation committee interested in resident uses of lakes described a strategy for finding that type of information: “So when we start to look out and search the Internet we throw a broad net at the beginning based on certain things like those lake management plans but when we get down to specifics we start looking at information of lakes that are more in the same latitude or in close proximity to where the municipality that we live is.” Another interviewee who worked with a local land trust took a

different approach: “In terms of my work and the way I would do it, it would be place based; it would be coming from the place to the information.”

In either case, interviewees found being able to begin their searches either by topic or place keywords was important for their search strategies.

5.6.2.2. Quantitative Results. Questionnaire respondents were asked how important the “Ability to search for data in different ways, e.g., by location, keyword, etc.” would be to them. The results are consistent with those from the interview phase of this research.

Figure 10. Importance of being able to search for data in different ways (n=139)



Being able to conduct searches using different starting points, including location and natural language keywords, appears to be an important functional capability for an online repository for locally generated, spatially-referenced data.

5.6.3. Hypothesis Subpart (c): Peer Evaluation

Data users would be willing to consider using data in an online data commons environment if such an environment included a simple post-publication peer-evaluation mechanism that would both provide feedback for contributors, and provide information on quality and suitability for use for users.

5.6.3.1. Qualitative Findings. In this age of Amazon and online shopping, it is no surprise that interviewees used online shopping comments as an analog to looking at comments/evaluations in an online commons environment for spatially referenced data. Half of the interviewees made comments similar to this one: “I mean I buy CDs on Amazon.com” that indicated familiarity with commercial online retailer commenting systems that they found useful, and indicating that they would consult peer comments and evaluation of data files if such comments were available.

Half of the respondents, however, said that they would look at the data themselves if it were data that might suit their needs, no matter what the comments said. Two indicated that they would look at the data first and only subsequently consult other user comments to see if those corresponded with their own judgments.

Only one interviewee said that the interviewee would be unlikely to consult comments made by others because the interviewee preferred to form a personal opinion directly from the data.

One interviewee indicated that “junk comments” were always a potential problem in evaluation systems and recommended that any such system have a

moderator who would screen comments for civility, relevance, and, if possible, quality before posting them.

Other interviewees who would consult comments made by others indicated that while they would not view it as necessary, they would prefer to know who the commenter was so that they could form an opinion about the relevance or quality of the comment source if the commenter were known to them.

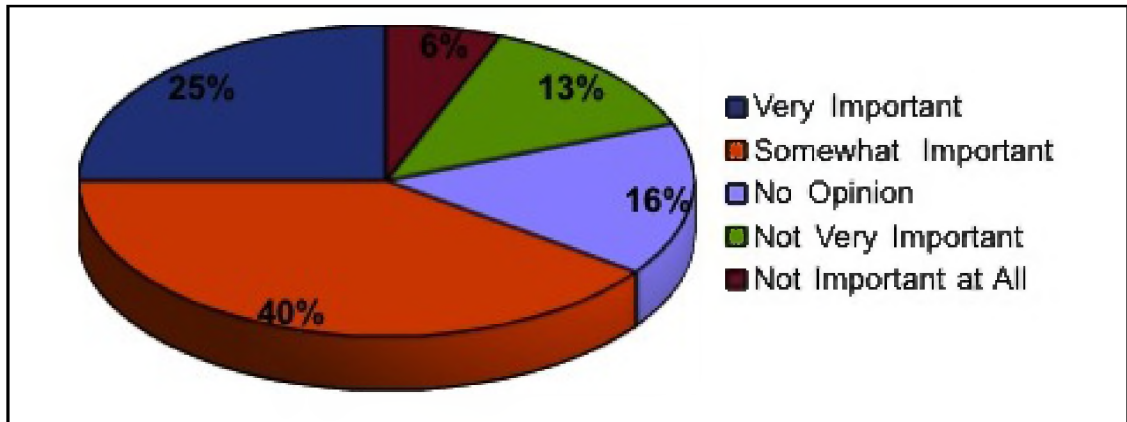
Nine of the interviewees indicated that they would be willing to make comments if they felt that they had something useful to say about a file. Most said that they would be willing to spend a limited amount of time, 5 to 15 minutes, to input a comment if there were a simple way to do so.

Consistent with the desire to know who made a comment, all nine said that they would be willing to use their own names rather than to use a screen name in offering a comment.

In summary, the majority of interviewees would find a commenting/ evaluation system valuable in an online commons repository.

5.6.3.2. Quantitative Results. Support for the “Ability to comment on the suitability of the data for your uses” was not so strong among survey respondents as among interviewees, although it was substantial, with 65 percent finding that capability “Very Important” or “Somewhat Important.”

Figure 11. Importance of being able to comment on suitability of data for use

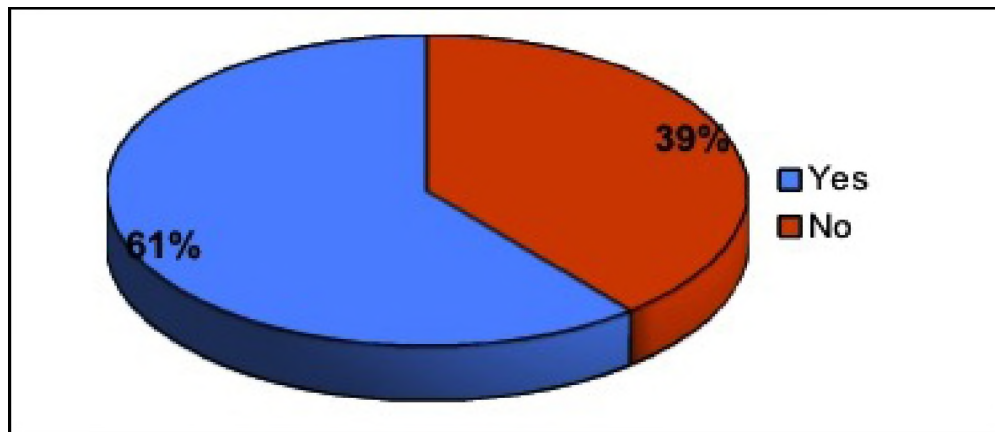


The amount of time that survey respondents would be willing to spend providing a comment generally mirrored what most interviewees would spend, 5 to 15 minutes. Given 139 responses rather than 10 as in the personal interviews, however, it is not surprising that there were a few outliers who would commit anywhere from “no time” to “as much as would be needed.”

In response to the question “Would the comments of other users affect your decision about whether to examine data that is available in the repository?” of 138 responses, 61 percent replied “Yes” and 39 percent said “No.”

Figure 12. Would comments of others affect your decision to examine data?

(n=138)

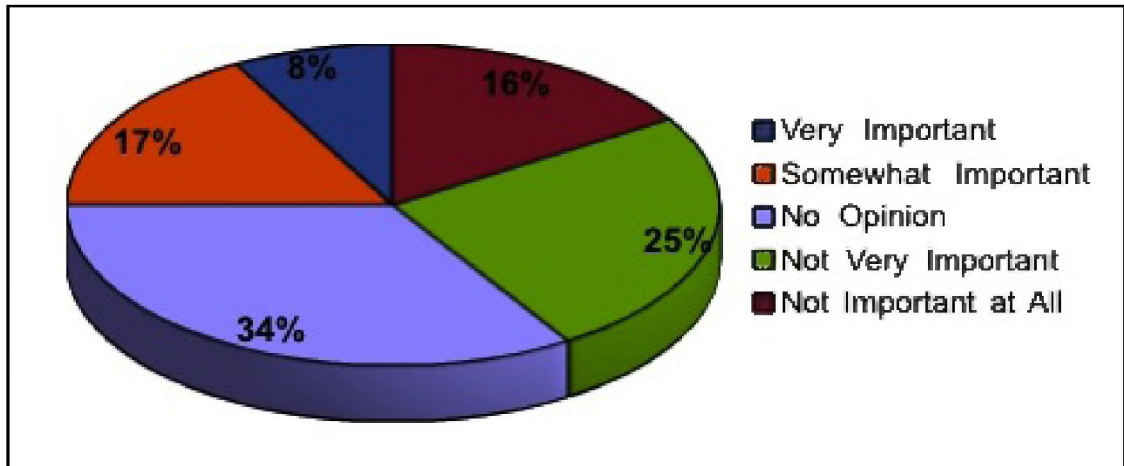


When asked to “explain how comments of others might affect your decision about whether to examine data further,” a large majority of those who answered (78 of 84) cited comments that dealt with data quality and accuracy. Here, again, the analogy of online commerce sites came up: “Same as eBay. If someone says the data are junk, I’ll probably be reluctant to use them.”

The other major reason expressed by respondents was not the quality of the data itself but rather the lack of suitability for purpose, e.g., “how the data fits with my base maps.”

The “Ability to use a screen name rather than your actual name when commenting” was more of an issue to survey respondents than it was with the interviewees.

Figure 13. Importance of using a screen name when commenting on data
(n=139)



While nine of ten interviewees would use their own names rather than a screen name when making comments and preferred to know the identity of those making comments when possible, 25 percent of questionnaire respondents felt it would be “Very Important” (8 percent) or “Somewhat Important” (17 percent) to be able use screen names when commenting, and a third did not express any opinion. The reason for this divergence from the attitudes of interviewees is not explainable based on the data this research gathered. The location of the questionnaire respondents might be an issue for commenting using one’s real name, or employment status, or some other variable for which this research did not gather any data.

5.7. Chapter Summary and Conclusions

This research, subject to the caveats listed below, empirically suggests that it would be desirable from the perspective of potential users of spatially referenced data in an online commons-type environment to provide infrastructure capability that would:

- make conditions of use of files clear to potential users,
- provide a variety of ways to search for data, and
- enable users to access comments and feedback from prior users, and to add comments of their own.

There are other desirable features of a commons-type online infrastructure, as the reports cited previously outline. This research addressed only these three.

5.7.1. Limitations

As noted earlier, this research has several limitations that prevent any assertion that the hypothesis is “proven” in the usual meaning of that term. However, we can assert that the hypothesis is supported by the results of this study.

These limitations do not, we feel, limit the usefulness of the research results for their intended purpose: to provide guidance to those who may in the future choose to construct an online commons environment for locally-generated, spatially-referenced data that anyone, nonprofessional and professional alike, can use.

5.7.2. Directions for Future Research

This research is based on interviews and on online questionnaire results. Results from the interviews generally are confirmed by the survey results. Although percentages differed slightly, opinions about the hypotheses generally were shared both in the interviews and in the survey responses.

However, there was a noticeable disparity in the perception of the importance of being able to use a screen name rather than a real name to make comments, although because a large number of questionnaire respondents expressed “No Opinion,” it is difficult to tell if the disparity was important. The absence of demographic, employment, or geographic location information for interviewees and questionnaire respondents makes it impossible to explain that divergence based on those characteristics. This is an area in which additional research may be fruitful.

This study made no effort to directly ask comparative questions, e.g., is one factor, such as clarity of conditions, more important than another to respondents? Answers to such questions may be inferred from the responses in the importance respondents placed on each factor, but it also could be desirable to ask comparative questions directly.

5.7.3. Possible Wider Applications

While this research focused on a possible future online commons-type environment for spatially referenced data from widely disparate sources, the results could be of some use to operators of existing online spatial-data services. Understanding what is desirable to users in approaching data with which they

are not familiar, especially non-GIS professionals, could be helpful for existing services to, for example, make clear in an obvious way any restrictions on use of their data. Portals that do not presently enable users to search for data in different ways may wish to evaluate whether such functionality would be desirable to their existing user base, and whether it might help to increase usage among current nonusers of their services. Sites that do not offer commenting capability may wish to investigate if that functionality might increase usage.

For designers of potential future online environments for spatially referenced data, which might include, for example, university libraries or state library systems, and possibly for operators of existing portals as well, we hope this research, though not designed to be statistically “proven,” offers some empirical insight into what online characteristics users find valuable for spatially referenced data repositories and/or portals.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1. Study Motivation

This study arose from a desire to answer a core question about the feasibility of constructing a Commons of Geographic Data, a question that had been raised by grant reviewers as well as others: if such a commons repository were built, would potential users and potential contributors be willing to use it, and, if so, under what conditions?

The question involves a number of different dimensions that are anchored in law and policy, as well as in technical domains, and therefore is a fitting question to explore within the context of the law and policy area concentration within the Spatial Informatics Program in the School of Computing and Information Science at the University of Maine.

The product of the research would help to not only answer whether potential users and contributors would be willing to use such a repository but would also provide useful information about desired functionality to information architects who might wish to construct a Commons of Geographic Data in the future, and could be of use to other repositories of spatially referenced and other types of scientific data.

6.2. Study Goals

This study sought to examine in Chapters 2 and 3 why a commons type repository for locally generated, spatially referenced data had a place in the larger universe of access to scientific data in today's digital environment. We looked at the policy, legal, and scientific contexts that such a repository, which we refer to as a Commons of Geographic Data, would fit into.

Within those contexts, we posited that potential contributors to such a repository would find three infrastructure functions desirable when considering whether to donate their data:

- allow users to attach conditions to the use of their data,
- provide basic information that could be translated into standards based metadata, and
- receive comments and feedback from users.

We also posited that potential users of such a repository would find three infrastructure functions desirable when considering whether to donate their data. These functions mirror those that contributors would find desirable but as seen from a user perspective:

- make conditions of use of files clear to potential users
- provide a variety of ways to search for data
- enable users to access comments and feedback from prior users, and to add comments of their own.

While some might assume that these hypotheses could be considered obvious and would hold, absent empirical verification, they would remain assumptions

only. In Chapters 4-5, we reported on the results of a combination of qualitative and quantitative research that addressed those posited hypotheses.

6.3. Conclusions Based on This Study

At present, there is no repository online for locally generated spatially referenced data which includes all of the desirable functionality described in Chapter 3. The research undertaken in this study indicates that the study hypotheses hold, and that both potential contributors (reported in Chapter 4) and potential users (reported in Chapter 5) would be interested in utilizing such a repository if it were built and included the posited functionality.

6.4. Recommendations for Information Architects of a Commons of Geographic Data

Based upon the results of this study, we would recommend that information architects who might undertake the design and construction of a Commons of Geographic Data or similar online repository include the following functionality in the site design:

- a. an ability for data contributors to easily indicate whether they want to put any conditions of the use of their data by others, If they do, a simple way to indicate whether they wish to have attribution if others use their data, whether they want their data restricted to non-commercial use, and whether they want to

restrict their data to being used only as is and not modified in any way for subsequent use. These are all options available under Creative Commons licenses

b. a way for data contributors to have the option to use natural language text and keyword descriptions of the contents of their contributed files that will render those descriptions as standards-based metadata without additional contributor effort; and/or to use controlled vocabularies

c. keeping the time required for the registering of a contributed file to 15 minutes or less

d. assuring potential contributors that their data will be maintained for a specific amount of time on the site

e. a peer commenting/evaluation system which enables users to review the comments of others about the usefulness of a data file for a particular purpose, and contribute their own comments, and allow those who wish to use a screen name other than their own when posting comments do so providing there is an actual person with a confirmed email address using the screen alias

f. identify conditions of use to potential users before files are viewed or downloaded

g. enable searching of the stored data files using a variety of search strategies including location, controlled vocabulary terms, natural language terms, and average comment ratings.

It is the author's hope that the question of "if you build it, will they come" has been answered sufficiently so that this will no longer be an issue in deciding

whether contributors or users would consider utilizing a Commons of Geographic data or similar repository.

The likelihood of maximizing use of such a repository would be enhanced by designing it to operate simply enough that non-professionals could use it comfortably, and to include the functionality listed above.

Perhaps we will one day see a Commons of Geographic Data built.

REFERENCES

- Agres, Ted. 2006. Tying up Science: Are Intellectual Property Protections Slowing Progress? *The Scientist* 20 (1): 77.
<http://www.thescientist.com/article/display/18850/>.
- Anderson, Chris. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired* 16 (7):
http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.
- Anderson, Melinda J., and Robin N. Shaw. 1999. A Comparative Evaluation of Qualitative Data Analytic Techniques in Identifying Volunteer Motivation in Tourism. *Tourism Management* 20 99-106.
- Anthes, Gary. 2009. Deep Data Dives Discover Natural Laws. *Communications of the ACM* 52 (11): 13-14. <http://cacm.acm.org/magazines/2009/11/48443-deep-data-dives-discover-natural-laws/fulltext>.
- ARL Workshop on New Collaborative Relationships: the Role of Academic Libraries in the Digital Data Universe, Amy Friedlander, Prudence Adler, and National Science Foundation (U.S.). 2006. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*. Washington: Association of Research Libraries.
- Arms, William Y, and Ronald L Larson. 2007. *The Future of Scholarly Communication: Building the Infrastructure of Cyberscholarship*. Washington: National Science Foundation.
- Armstrong, Timothy K. 2006. Digital Rights Management and the Process of Fair Use. *Harvard Journal of Law and Technology* 20 (1): 49-121.
- Armstrong, Timothy K. 2008. Fair Circumvention. *Brooklyn Law Review* 74 (1): 1-50.
- Asay, Clark D. 2013. Kirtsaeng and the First-Sale Doctrine's Digital Problem. *Stanford Law Review* 66
<http://www.stanfordlawreview.org/online/kirtsaeng-and-first-sale-doctrines-digital-problem>.
- Ball, Alex, and Monica Duke. 2012. How to Cite Datasets and Link to Publications: a Report of the Digital Curation Centre.
<http://codata2012.tw/sites/default/files/text/slide/codata2012-Duke%20and%20Ball-How%20to%20Cite%20Datasets%20and%20Link%20to%20Publications.pdf>.

- Barabasi, Albert-Laszlo. 2002. *Linked: The New Science of Networks*. Cambridge, MA: Perseus Publishing.
- Bishr, Mohamed, and Lefteris Mantelas. 2008. A Trust and Reputation Model for Filtering and Classifying Knowledge About Urban Growth. *GeoJournal* 72 229-37. DOI 10.1007/s10708-008-9182-4.
- Boldrin, Michele, and David K. Levine. 2002. The Case Against Intellectual Property. *American Economic Review Papers and Proceedings* 92 209-12.
- Boseley, Sarah. 2009. Drug Giant GlaxoSmithKline Pledges Cheap Medicine for World's Poor. <http://www.theguardian.com/business/2009/feb/13/glaxo-smith-kline-cheap-medicine>.
- Boyle, James. 2003. The Second Enclosure Movement and the Construction of the Public Commons. *Law and Contemporary Problems* 66 (33): 33-75.
- Breyer, Stephen. 1970. The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs. *Harvard Law Review* 84 (2): 281-351.
- Brindley, Lynne. 2007. Balance in IP "Not working". http://legacy.earlham.edu/~peters/fof/2007_12_02_fosblogarchive.html.
- Bruns, Axel. 2007. *Producers: Towards a Broader Framework for User-led Content Creation*. In *Cc 2007: Creativity & Cognition 2007: Seeding Creativity—tools, Media, and Environments*, edited by Ben Shneiderman. New York: Association for Computing Machinery.
- Campbell, James, Marilyn Lutz, David McCurry, Harlan Onsrud, and Kenton Williams. 2006. *Enabling Non-specialist Contributors to Generate Standards-based Geographic Metadata in a Commons of Geographic Data*. In *Abstract Proceedings of Giscience 2006*, Muenster: Institute for Geoinformatics.
- Caplan, Priscilla, Bill Barnett, Liz Bishoff, Christine Borgman, Ken Hamma, and Clifford Lynch. 2003. Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content. <http://www.ims.gov/pdf/digitalopp.pdf>.
- Carlson, Scott. 2006. Lost in a Sea of Science Data. *Chronicle of Higher Education* 52 (42): <http://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136>.
- CODATA/ICSU. 2014. Big Data for International Scientific Programmes: Challenges and Opportunities. Paper read at Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities, June 8-9, at Beijing.

- Committee on Science, Engineering, and Public Policy (U.S.). 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington: National Academies Press.
- Cooper, Seth, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic, and Foldit Aplayers. 2010. Predicting Protein Structures with a Multiplayer Online Game. *Nature* 466 (7307): 756-60. <http://dx.doi.org/10.1038/nature09304>.
- Creative Commons. 2014. Data. <http://wiki.creativecommons.org/Data>.
- Creswell, John W. 2007. *Qualitative Inquiry and Research Design: Choosing Among Five Traditions*. Thousand Oaks: Sage Publications.
- Eschenfelder, Kristin R, and Michelle Caswell. 2010. Digital Cultural Collections in an Age of Reuse and Remixes. *First Monday* 15 (11): <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/Article/3060/260>.
- European Commission. 2013. Guidelines on Data Management in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.
- European Commission. 2013. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.
- Falzone, Anthony. 2009. URAA Held Unconstitutional. <http://cyberlaw.stanford.edu/node/6149>.
- Flanagin, Andrew J., and Miriam J. Metzger. 2008. The Credibility of Volunteered Geographic Information. *GeoJournal* 72 137-48. DOI 10.1007/s10708-008-9188-y.
- Gartner, G, D Bennett, and T Morita. Toward Ubiquitous Cartography. *Cartography and Geographic Information Science* 34 247-57.
- GEOSS. 2010. GEOSS Data Sharing Action Plan. https://www.earthobservations.org/documents/geo_vii/07_GEOSS%20Data%20Sharing%20Action%20Plan%20Rev2.pdf.
- Goodchild, Michael F. 2007. Citizens As Sensors: The World of Volunteered Geography. *GeoJournal* 69 211-21.
- GovLab. 2014. Open Data 500. <http://www.opendata500.com>.

- Hars, Alexander, and Shaosong Ou. 2002. Working for Free? Motivations for Participating in Open-Source Projects. *International Journal of Electronic Commerce* 6 (3): 25-39.
- Hearst, Marti, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. 2002. Finding the Flow in Web Site Search. *Communications of the ACM* 45 (9): 42-49.
- Herper, Matthew, and Robert Langreth. 2007. Biology Goes Open Source. http://www.forbes.com/2007/02/12/novartis-genes-diabetes-research-biz-cz_mh_0212novartis.html?partner=rss.
- Heverly, Robert A. 2003. The Information Semicommons. *BTLJ* 18 1127-89.
- Hohfeld, Wesley Newcomb. 1978. *Fundamental Legal Conceptions As Applied in Judicial Reasoning*. Edited by Cook, Walter Wheeler. Westport, CT: Greenwood Press.
- Institute of Museum and Library Services (U.S.). 2003. *Report of the Workshop on Opportunities for Research on the Creation, Management, Preservation and Use of Digital Content*. Washington: Institute of Museum and Library Services.
- Interagency Working Group on Digital Data. 2009. *Harnessing the Power of Digital Data for Science and Society*. Washington: National Science and Technology Council, Executive Office of the President.
- International Publishers Association. 2014. Annual Report 2013-2014. <http://www.internationalpublishers.org/images/reports/2014/IPA-annual-report-2014.pdf>.
- Kansa, Eric C., Jason Schultz, and Ahrash N. Bissell. 2005. Juxtaposing Intellectual Property Agendas via a "Some Rights Reserved" Model. *International Journal of Cultural Property* 12 (3): 285-314.
- Kash, Wyatt. 2014. Why Free Government Data Remains a Tough Sell. *Information Week* <http://www.informationweek.com/government/open-government/why-free-government-data-remains-a-tough-sell/d/d-id/1113604>.
- Kelsey, John, and Bruce Schneier. 1999. The Street Performer Protocol and Digital Rights. *First Monday* 4 (6-7): <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/673/583>.
- King, Bob. 2012. Rick Santorum's Campaign Could Be Clouded by 7-year Old Attack on National Weather Service. <http://www.politico.com/news/stories/0112/71129.html>.

- Lakhani, Karim R., and Robert G. Wolf. 2005. *Why Hackers Do What They Do: Understanding Motivation and Effort in Free/open Source Software Projects*. In *Perspectives on Free and Open Source Software*, edited by J. Feller, B. Fitzgerald, S. Hissam, and K. R. Lakhani. Cambridge, MA: MIT Press.
- Lemley, Mark. 2004. Property, Intellectual Property, and Free Riding. <http://ssrn.com/abstract=582602>.
- Lessig, Lawrence. 2003. The Future of Ideas and Code and Other Laws of Cyberspace. Paper read at 13th Annual Conference on Computers, Freedom & Privacy, Apr 1-4, at New York.
- Lessig, Lawrence. 2004. Building the Creative Commons. <http://www.uwm.edu/Dept/SOIS/about/news/events/index.htm>.
- Lessig, Lawrence. 2004. *Free Culture: How Big Media Uses Technology and the Law to Lock Down Culture and Control Creativity*. New York: Penguin Press.
- Lipton, Jacqueline. 2004. Information Property: Rights and Responsibilities. *Fla. L. Rev.* 56 (1): 140.
- LOCKSS. 2010. What Is the LOCKSS Program? <http://www.lockss.org/about/what-is-lockss>.
- Lunney, Glynn S., Jr. 1996. Reexamining Copyright's Incentives-Access Paradigm. *Vanderbilt Law Review* 49 483-656.
- Mathiesen, Kay. 2009. Access to Information as a Human Right. Paper read at iConference 2009, Feb. 8-11, at Chapel Hill.
- Maxwell, Joseph A. 2005. *Qualitative Research Design: An Interpretive Approach*. 2nd ed. Vol. 41, *Applied Social Research Series*. Thousand Oaks, CA: Sage Publications.
- McCarty, L. Thorne. 2002. Ownership: A Case Study in the Representation of Legal Concepts. *Artificial Intelligence and Law* 10 (1-3):
- McKee, Lance. 2010. Seventeen reasons why geospatial research data should be published online using OGC standard interfaces and ISO Standard Metadata. <http://blog.okfn.org/2010/06/21/open-geoprocessing-standards-and-open-geospatial-data/>.
- McKenzie, Pamela J, Jacquelyn Burkell, Lola Wong, Caroline Whippley, Samuel E Troscow, and Michael McNally. 2012. User-generated Online Content 1: Overview, Current State and Context. *First Monday* 17 (6): <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/Article/3912/326>.

- Murillo, Angela P. 2013. Data at Risk Initiative: Examining and Facilitating the Scientific Process in Relation to Endangered Data. *Data Science Journal* 12 207-19. <http://dx.doi.org/10.2481/dsj.12-048>.
- Murray-Rust, Peter. 2007. Data-driven Science - a Scientist's View. Paper read at NSF/JISC Repositories Workshop, Apr 17-18, at Phoenix.
- National Information Standards Organization, and National Federation of Advanced Information Services. 2013. Recommended Practices for Online Supplemental Journal Article Materials. http://www.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf.
- National Research Council (U.S.) Committee on Licensing Geographic Data and Services. 2004. *Licensing Geographic Data and Services*. Washington: National Academies Press.
- National Science Board. 2005. *Long-Lived Digital Data Collections: Research and Education in the 21st Century*. Washington: National Science Foundation.
- National Science Foundation (U.S.). 2011. Grant Proposal Guide (Chapter II). http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp.
- National Science Foundation Cyberinfrastructure Council. 2006. *NSF's Cyberinfrastructure Vision for the 21st Century (v5.0)*. Washington: National Science Foundation.
- New York Law School Institute for Information Law and Policy. 2009. *Access to Health Information Under International Human Rights Law (Draft)*. New York: New York Law School.
- Nielsen, Jakob. 1993. *Usability Engineering*. Boston: Academic Press.
- Nielsen, Jakob. 2000. *Designing Web Usability*. Indianapolis IN: New Riders.
- Nov, Obed. 2007. What Motivates Wikipedians? *Communications of the ACM* 50 (11): 60-64.
- ÓhAnluain, Daithi. 2004. Calls for Open Access Challenge Academic Journals. <http://www.ojr.org:80/ojr/stories/121004ohanluain/>.
- Olson, Hope A. 1998. Mapping Beyond Dewey's Boundaries: Constructing Classificatory Space for Marginalized Knowledge Domains. *Library Trends* 47 (2): 233-54.

- Onsrud, Harlan, Gilberto Camara, James Campbell, and Narindi Sharad Chakravarthy. 2004. *Public Commons of Geographic Data: Research and Development Challenges*. In *Geographic Information Science*, edited by Max J. Egenhofer, Christian Freska, and Harvey J. Miller. Berlin: Springer-Verlag.
- Onsrud, Harlan, and James Campbell. 2007. Big Opportunities in Access to "Small Science" Data. *Data Science Journal* 6.
- Onsrud, Harlan, James Campbell, and Bastiaan van Loenen. 2010. Towards Voluntary Interoperable Open Access Licenses for the Global Earth Observation System of Systems (GEOSS). *International Journal of Spatial Data Infrastructures Research* 5 194-215.
<http://ijmdir.jrc.ec.europa.eu/index.php/ijmdir/article/view/168>.
- Onwuegbuzie, Anthony J., and Nancy L. Leech. 2004. Enhancing the Interpretation of "Significant" Findings: The Role of Mixed Methods Research. *The Qualitative Report* 9 4 770-92.
- Pallante, Maria A. 2013. The Next Great Copyright Act. *Columbia Journal of Law and the Arts* 36 (3): 315-44.
http://www.copyright.gov/docs/next_great_copyright_act.pdf.
- Parchomovsky, G, and P.J. Weiser. 2010. Beyond Fair Use. *Cornell Law Review* 96 (1): 91-137.
- Pessach, Guy. 2008. Reciprocal Share-Alike Exemptions in Copyright Law. *Cardozo Law Review* 30 (3): 101-50.
- Public Policy Committee of the ACM. 2006. USACM Policy Recommendations on Digital Rights Management. <https://freedom-to-tinker.com/blog/felten/usacm-policy-statement-drm/>.
- Ragin, Charles, Joane Nagel, and Patricia White. 2004. *Workshop on Scientific Foundations of Qualitative Research*. Washington, DC: National Science Foundation.
- Registrar of Copyrights. 2006. *Report on Orphan Works*. Washington: United States Copyright Office.
- Reichman, Jerome H, Graeme B Dinwoodie, and Pamela Samuelson. 2007. A Reverse Notice and Takedown Regime to Enable Public Interest Use of Technically Protected Copyrighted Works. *Berkeley Technology Law Journal* 22 981-1060.
- Riemer, Harold A., Kim D. Dorsch, Larena Hoeber, David Paskevich, and Packianathan Chelladurai. 2004. *Motivations for Volunteering with Youth-Oriented Programs*. Toronto: Canadian Centre for Philanthropy.

- Robinson, W.S. 1951. The Logical Structure of Analytic Induction. *American Sociological Review* 16 (6): 812-18.
- Royal Society of Chemistry. 2014. ChemSpider Terms and Conditions. <http://www.rsc.org/help/termsconditions.asp>. .
- Samuelson, Pamela. 2003. DRM {and, or, vs.} the Law. *Communications of the ACM* 46 (4): 41-45.
- Samuelson, Pamela. 2007. Preliminary Thoughts on Copyright Reform. *Utah Law Review* 2007 (3): 551-71.
- Scholarly Publishing and Academic Resources Coalition (SPARC). 2006. What Is SPARC? <http://www.arl.org/sparc/about/index.html>.
- Shaver, Lea. 2010. The Right to Science and Culture. *Wisconsin Law Review* 2010 (1): 121-84.
- Shirky, Clay. 2003. Power Laws, Weblogs, and Inequality. http://www.shirky.com/writings/powerlaw_weblog.html. Accessed on Jan 12, 2006.
- Sohn, Gigi B. 2007. Six Steps to Digital Copyright Sanity: Reforming a Pre-vcr Law for a Youtube World. Paper read at New Media and Marketplace of Ideas Conference, Oct. 26, at Boston.
- Sprigman, Christopher. 2004. Reform(aliz)ing Copyright. *Stanford Law Review* 56 (2): 485-568.
- Sui, Daniel Z. 2008. The Wikification of GIS and its Consequences: Or Angelina Jolie's new tattoo and the Future of GIS. *Computers, Environment and Urban Systems* 32 (1): 1-5.
- Szalay, Alexander, and Jim Gray. 2006. 2020 Computing: Science in an Exponential World. *Nature* 440 413-14. <http://www.nature.com/nature/journal/v440/n7083/full/440413a.html>.
- The Royal Society. 2003. Keeping Science Open: the Effects of Intellectual Property Policy on the Conduct of Science. <http://www.royalsoc.ac.uk/document.asp?tip=0&id=1374>.
- Turner, Andrew. 2006. *Introduction to Neogeography*.: O'Reilly Media, Inc.
- Tuutti, Camille. 2011. NSF Seeks Cyber Infrastructure to Make Sense of Scientific Data. <http://fcw.com/Articles/2011/10/04/NSF-taps-UNC-researchers.aspx?Page=1>.

- United Nations. 1948. Universal Declaration of Human Rights.
<http://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=eng>.
- University of California. 2014. DMPTool. <https://dmp.cdlib.org>.
- van der Hoeven, Jeffrey, Salvatore Mele, Veronica Guidetti, and Sabine Schrimpf. 2010. What We Learned From Parse.insight. Paper read at PARSE.Inisght Symposium, June 2010, at Paris, France.
- Weber, Steven. 2004. *The Success of Open Source*. Cambridge MA: Harvard University Press.
- Whitlock, Michael C., Mark A. McPeck, Mark D. Rausher, Loren Rieseberg, and Allen J. Moore. 2010. Data Archiving. *The American Naturalist* 175 (2): 145-46.
- Yale University. 2014. Yale University Open Data Access (YODA) Project.
<http://medicine.yale.edu/core/projects/yodap/index.aspx>.

APPENDIX: Table 2. Data repository sites with more complete description of usage rights.

Site	Description	url	Usage information
Scientific Earth Drilling Information Service - SEDIS	The Integrated Ocean Drilling Program (IODP) is developing a web based information service SEDIS - to facilitate access to all data and information related to scientific ocean drilling, regardless of origin or location of data. SEDIS will be designed to integrate distributed scientific drilling data via metadata.	http://sedis.iodp.org/front_content.php	No mention of usage rights. Data sets can be downloaded right from site
Data.gov	Data.gov is the official portal for open data from the U.S. government. It is a public domain website	http://www.data.gov	U.S. Federal data available through Data.gov is offered free and without restriction. Data and content created by government employees within the scope of their employment are not subject to domestic copyright protection under 17 U.S.C. § 105. Non-federal data available through Data.gov may have a different licensing method as noted under "Show more" at the bottom of the dataset page. Non-federal data can be identified by name of the publisher and the diagonal banner that shows up on the search results and data set pages. Federal data will have a banner noting "Federal" and non-federal banners will note "University", "Multiple Sources", "State", etc."

PubChem	PubChem, released in 2004, provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative.	http://pubchem.ncbi.nlm.nih.gov/	"Information that is created by or for the US government on this site is within the public domain... This site contains resources such as, but not limited to, PubMed Central (see PMC Copyright Notice), Bookshelf (see Bookshelf Copyright Notice), OMIM, and PubChem which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws."
Online Mendelian Inheritance in Man (OMIM®)	OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available and updated daily.	http://www.omim.org/help/copyright	"The rights in and to OMIM (excluding information contained therein obtained from third parties) vest in JHU. JHU holds the copyright and trademark to OMIM and OMIM.org, including the collective data therein... Use of OMIM.org is provided free of charge to any individual for personal use, for educational or scholarly use, or for research purposes through the front end of the database."
Montana Geographic Information Clearinghouse		http://geoinfo.rsl.mt.gov/	No usage information stated

MetroGis	The purpose of MetroGIS is to institutionalize the sharing of accurate and reliable geospatial data so user and producer communities can share in the efficiencies of being able to effortlessly obtain the data they need, in the form they need, when they need it.	http://metrogis.org/	“...government data are public and are accessible by the public for both inspection and copying unless there is federal law, a state statute, or a temporary classification of data that provides that certain data are not public.”
BOLD	The Barcode of Life Data Systems (BOLD) is an informatics workbench aiding the acquisition, storage, analysis, and publication of DNA barcode records. By assembling molecular, morphological, and distributional data, it bridges a traditional bioinformatics chasm. BOLD is freely available to any researcher with interests in DNA barcoding.	http://www.barcodinglife.org/views/login.php	Incorporates data from GenBank, Canadian Centre, others. Makes what it refers to data as public data available for search or download but does not discuss usage or copyright

<p>ChemSpider [example of a site offering access but not re-use]</p>	<p>ChemSpider is a free chemical structure database providing fast access to over 30 million structures, properties and associated information, and 400 data sources, ChemSpider enables researchers to discover the most comprehensive view of freely available chemical data from a single online search. It is owned by the Royal Society of Chemistry.</p>	<p>http://www.chemspider.com/</p>	<p>You may browse, download or print out one copy of the material displayed on the site for your personal, non-commercial, non-public use, but you must retain all copyright and other proprietary notices contained on the materials. You may not further copy, distribute or otherwise use any of the materials from this site without the advance, written consent of RSC.</p>
--	--	--	---

<p>Freebase</p>	<p>Initially, Freebase was seeded by pulling in information from a large number of high-quality open data sources, such as Wikipedia, MusicBrainz, and others. The Freebase community along with the internal Freebase team continue to drive the growth of the graph by focusing on bulk, algorithmic data imports, data extraction from free text, ongoing synchronization of data feeds, and rigorous quality management.</p>	<p>http://www.freebase.com/</p>	<p>CC-By license and some under GFDL</p>
-----------------	--	--	--

Sage Bionetworks	We work to redefine how complex biological data is gathered, shared and used, redefining it through open systems, incentives, and norms.	http://sagebase.org/	Our software is available in Github, and our non-software creative works are licensed under the Creative Commons Attribution 3.0 Unported license except for legacy publications in closed journals. The research projects benefit both the specific collaborators and the larger scientific community because the results will also be accessible in the Sage Bionetworks Commons one year after the conclusion of the research projects.
uBio	Indexing & Organizing 11,106,374 Biological Names. uBio is an initiative within the science library community to join international efforts to create known names of all living (and once-living) organisms and utilize a comprehensive and collaborative catalog of known names of all living (and once-living) organisms.	http://www.ubio.org/	Many tools and applications. No specific rights info.

ICDNS	<p>To date criteria have been developed by essentially closed groups of interested workers and this may have limited the speed of development and responsiveness of classification schemes.</p> <p>To make such criteria widely accepted many people now believe that there should be an opportunity for any interested worker to participate in their development. Such a democratic forum can now be realised using the internet and the web.</p>	<p>http://www.icdns.org/</p>	<p>“Use, reproduction and intellectual property in the contents of the ICDNS website are assigned according to an 'Open Source' license agreement which is presented in the Discussion Forum. This allows free use of material provided that the user complies with the terms of the license. You must AGREE to the terms of this license before making any use of material on the website.” [THIS PAGE NOT AVAILABLE AS OF 7/1/14]</p>
ZooBank	<p>ZooBank provides a means to register new nomenclatural acts, published works, and authors.</p>	<p>http://www.zoobank.org/</p>	<p>Rights usage not specifically noted but see paper on Scientific names of organisms</p>
OneGeology [U.S. not a member but federal and state agencies make data available]	<p>OneGeology's aim is to create dynamic digital geological map data for the world. It is an international initiative of the geological surveys of the world who are working together to achieve this ambitious and exciting venture.</p>	<p>http://www.onegeology.org</p>	<p>“Map data distributed as part of OneGeology will remain in the ownership of the originating geological survey or organisation, and ideally be available at no cost.”</p>

DOE Data Explorer	Use the DOE Data Explorer (DDE) to find scientific research data - such as computer simulations, numeric data files, figures and plots, interactive maps, multimedia, and scientific images - generated in the course of DOE-sponsored research in various science disciplines.	http://www.osti.gov/dataexplorer/	Public domain but... “When using the OSTI website, you may encounter documents, illustrations, photographs, or other information resources contributed or licensed by private individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws. Transmission or reproduction of protected items beyond that allowed by <u>fair use</u> as defined in the copyright laws requires the written permission of the copyright owners.”
NEXTBIO	NextBio is the provider of an innovative platform that enables life science researchers to search, discover, and share knowledge locked within public and proprietary data. NextBio's platform seamlessly combines powerful tools with unique correlated content to transform information into knowledge, providing the foundation for new scientific discoveries.	http://www.nextbio.com/	“NextBio contains the world's largest repository of curated correlated public and private genomic data, including data from multiple public repositories of genomic studies and patient molecular profiles, up-to-date reference genomes, and clinical trial results. Diverse molecular data types from these resources are systematically processed, curated and integrated into our private data center-based platform”

ChemBank	<p>ChemBank is a public, web-based informatics environment created by the Broad Institute's Chemical Biology Program and funded in large part by the National Cancer Institute's Initiative for Chemical Genetics (ICG). This knowledge environment includes freely available data derived from small molecules and small-molecule screens, and resources for studying the data so that biological and medical insights can be gained.</p>	<p>http://chembank.broadinstitute.org/</p>	<p>"The goals of ChemBank are to provide life scientists unfettered access to biomedically relevant data and tools heretofore available almost exclusively in the private sector. We intend for ChemBank to be a planning and discovery tool for chemists, biologists, and drug hunters anywhere, with the only necessities being a computer, access to the Internet, and a desire to extract knowledge from public experiments whose greatest value is likely to reside in their collective sum."</p>
LTSRF	<p>Long Term Stewardship and Reanalysis Facility (LTSRF) for the Group for High Resolution SST (GHRSSST), which is routinely delivering individual as well as multi-sensor blended SST products with high accuracy and fine spatial resolution</p>	<p>http://ghrsst.nodc.noaa.gov</p>	<p>National Oceanic Data Center. "NODC maintains the long term archive and works with the NASA JPL/Caltech Physical Oceanography Distributed Active Archive Center (PO.DAAC) Global Data Assembly Center (GDAC) to provide stewardship of these valuable data sets" [US Gov - public domain]</p>

ZINC	<p>Welcome to ZINC, a free database of commercially-available compounds for virtual screening. ZINC contains over 35 million purchasable compounds in ready-to-dock, 3D formats. ZINC is provided by the Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF).</p>	<p>http://zinc.docking.org/index.shtml</p>	<p>ZINC is freely available to everyone to use. Significant portions of ZINC may not be re-distributed without express written permission of John Irwin.</p>
------	---	--	--

GeoGratis	<p>GeoGratis is a portal provided by the <u>Earth Sciences Sector</u> (ESS) of Natural Resources Canada (NRCan) which provides geospatial data at no cost and without restrictions via your Web browser.</p>	<p>http://www.geogratias.cgdi.gc.ca</p>	<p>"Canada grants to the licensee a non-exclusive, fully paid, royalty-free right and licence to exercise all intellectual property rights in the data. This includes the right to use, incorporate, sublicense (with further right of sublicensing), modify, improve, further develop, and distribute the Data; and to manufacture or distribute derivative products." Attribution is required under Open Government Licence-Canada</p>
-----------	--	--	--

GBIF	<p>“The Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments. It allows anyone, anywhere to access data about all types of life on Earth, shared across national boundaries via the Internet.... It provides a single point of access (through this portal and its web services) to more than 400 million records, shared freely by hundreds of institutions worldwide, making it the biggest biodiversity database on the Internet.”</p>	<p>http://www.gbif.org</p>	<p>“The Participants who have signed the MoU have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data.</p> <p>GBIF data sharing should take place within a framework of due attribution.”</p>
<p>LinkedGeoData [although not hosted in the U.S., includes VGI data from U.S. contributors]</p>	<p>LinkedGeoData is an effort to add a spatial dimension to the Web of Data / Semantic Web. LinkedGeoData uses the information collected by the OpenStreetMap project and makes it available as an RDF knowledge base according to the Linked Data principles. It interlinks this data with other knowledge bases in the Linking Open Data initiative.</p>	<p>http://linked-geodata.org/About</p>	<p>The Linked Geo Data database is made available under the Open Database License. Any rights in individual contents of the database are licensed under the Database Contents License.</p>

dbGaP	<p>The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits</p>	<p>http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html</p>	<p>dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.</p>
Open Context	<p>Open Context is a free, open access resource for the electronic publication of primary field research from archaeology and related disciplines. It emerged as a means for scholars and students to easily find and reuse content created by others, which are key to advancing research and education. Open Context's technologies focus on ease of use, open licensing frameworks, informal data integration and, most importantly, data portability</p>	<p>http://opencontext.org/</p>	<p>"Open Context provides a platform for researchers to publish their primary field data and documentation. Because Open Context is a free and open access service, all members of the public are welcome to use and reuse this content."</p> <p>"Open Context licenses all content with Creative Commons, and makes it available in a variety of machine-readable formats."</p>

RRUFF	The RRUFF™ Project is creating a complete set of high quality spectral data from well characterized minerals and is developing the technology to share this information with the world. Our collected data provides a standard for mineralogists, geoscientists, gemologists and the general public for the identification of minerals both on earth and for planetary exploration.	http://rruff.info/	No specific rights info Appears to be OA – funded in part by NSF – also has private contributors.
PCL Map Collection	Maps digitized by the Univ. of Texas Libraries.	http://www.lib.utexas.edu/maps/	Most of the maps scanned by the University of Texas Libraries and served from this web site are in the public domain. A few maps are copyrighted, and are clearly marked as such.
ORegAnno [latest entry 2008]	AN OPEN ACCESS DATABASE FOR GENE REGULATORY ELEMENT AND POLYMORPHISM ANNOTATION The Open REGulatory ANNOTation database (ORegAnno) is an open database for the curation of known regulatory elements from scientific literature	http://www.oregano.org/oregano/Index.jsp	This project was funded by Genome Canada, the Michael Smith Foundation for Health Research, the Natural Sciences and Engineering Research Council, and the Canadian Institute for Health Research. It will receive ongoing maintenance and support from 2005 through 2007 [now listed in DataBib through Canada's Michael Smith Genome Sciences Centre

<p>antbase [latest entry appears to be 2009]</p>	<p>Antbase now provides for the first time access to all the ant species of the world, one of the ecologically most important groups of animals worldwide.</p>	<p>www.antbase.org</p>	<p>CC – By-NC-SA</p>
<p>AntWeb</p>	<p>AntWeb focuses on specimen level data and images linked to specimens. In addition, contributors can submit natural history information and field images that are linked directly to taxonomic names. Distribution maps and field guides are generated automatically. All data in AntWeb are downloadable by users. AntWeb also provides specimen-level data, images, and natural history content to the Global Biodiversity Information Facility (GBIF), the Encyclopedia of Life (EOL.org), and Wikipedia.</p>	<p>www.antweb.org</p>	<p>AntWeb content is licensed under a Creative Commons Attribution License. We encourage use of AntWeb images. In print, each image must include attribution to its photographer and "from www.AntWeb.org" in the figure caption. For websites, images must be clearly identified as coming from www.AntWeb.org, with a backward link to the respective source page. Photographer and other copyright information is provided on the big image page. Some photos and drawing belong to the indicated persons or organizations and have their own copyright statements. Photos and drawings with CCBY, CC-BY-NC or CC-BY-SA can be used without further permission, as long as guidelines above for attribution are followed.</p>

OpenStreetMap	OpenStreetMap is a free editable map of the whole world. It is made by people like you.	http://www.openstreetmap.org/	OpenStreetMap is open data, licensed under the Open Data Commons Open Database License (ODbL). The cartography in our map tiles, and our documentation, are licensed under the Creative Commons Attribution-ShareAlike license (CC BY-SA).
TOXNET [not listed in DataBIB]	Toxicology Data Network	http://toxnet.nlm.nih.gov/	Government information at NLM Web sites is in the public domain. Public domain information may be freely distributed and copied, but it is requested that in any subsequent use the National Library of Medicine (NLM) be given appropriate acknowledgement. When using NLM Web sites, you may encounter documents, illustrations, photographs, or other information resources contributed or licensed by private individuals, companies, or organizations that may be protected by U.S. and foreign copyright laws. Transmission or reproduction of protected items beyond that allowed by fair use as defined in the copyright laws requires the written permission of the copyright owners. Specific NLM Web sites containing protected information provide additional notification of conditions associated with its use.

<p>GlycomeDB [latest entry seems to be 2102 - copyright notice is 2007]</p>	<p>With this library we have translated the carbohydrate sequences of all freely available databases (CFG , KEGG, GLYCOSCIENCES.de, BCSDDB and Carbbank) to GlycoCT, and created a new database (GlycomeDB) containing all structures and annotations.</p>	<p>http://www.glycome-db.org/</p>	<p>Database of OA databases so presumably OA although there is a copyright notice on bottom of page</p>
---	--	--	---

<p>OBIS</p>	<p>OBIS (Ocean Biogeographic Information System) strives to document the ocean's diversity, distribution and abundance of life. Created by the Census of Marine Life, OBIS is now part of the Intergovernmental Oceanographic Commission of UNESCO, under its International Oceanographic Data and Information Exchange programme</p>	<p>http://www.iobis.org/home</p>	<p>OBIS is committed to keeping its data free and openly accessible for the public. So, if you have sensitive data you probably don't want to publish it through OBIS (or any other publication). OBIS does not claim ownership or rights to the data sets it publishes. All rights remain with the data source, whether distributed directly or mediated, whom may at any time decide to remove their data from OBIS</p>
-------------	---	--	---

ChEMBL	<p>The European Bioinformatics Institute is part of <u>EMBL</u>, Europe's flagship laboratory for the life sciences. EMBL-EBI provides freely available, covering the full spectrum of molecular biology. European Bioinformatics Institute - Funded by the Wellcome Trust</p>	<p>https://www.ebi.ac.uk/chembl/</p>	<p>Open - Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends on research consent agreements.</p>
GeoNames	<p>GeoNames contains over 10 million geographical names and consists of over 8 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes. The data is accessible free of charge through a number of webservice and a daily database export.</p>	<p>www.geonames.org</p>	<p>The GeoNames geographical database is available for download free of charge under a creative commons attribution license.</p>

Dryad	<p>The Dryad Digital Repository is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes.</p>	datadryad.org	<p>Repository Users are allowed and encouraged to reuse Content from the Repository in any manner except as described herein under "Prohibited Uses Generally" (Section 8.2) ["unlawful manner"]. To the extent possible under law, Submitters have waived all copyright and related or neighboring rights to this data.</p>
The National Map	<p>As one of the cornerstones of the U.S. Geological Survey's (USGS) National Geospatial Program, <i>The National Map</i> is a collaborative effort among the USGS and other Federal, State, and local partners to improve and deliver topographic information for the Nation.</p>	<p>http://nationalmap.gov</p>	<p>USGS-authored or produced data and information are considered to be in the U.S. public domain. While the content of most USGS Web pages is in the U.S. public domain, not all information, illustrations, or photographs on our site are. Some non USGS photographs, images, and/or graphics that appear on USGS Web sites are used by the USGS with permission from the copyright holder.</p>

BIOGRAPHY OF THE AUTHOR

James Campbell was born in Pennsylvania and grew up in Pennsylvania and New Jersey. He attended and graduated from high school in Jersey City, New Jersey. He attended LeMoyne College and received a BA in History, attended New York University and received an MA in American Civilization, and later attended the University of Wisconsin-Milwaukee and received a Masters in Library and Information Science degree. He has been self-employed during his work career. He is a candidate for the Doctor of Philosophy Degree in Spatial Information Science and Engineering from the University of Maine in May, 2015.