

## ABSTRACT

Title of Dissertation: DATA SHARING ACROSS RESEARCH AND PUBLIC COMMUNITIES

Yurong He, Doctor of Philosophy, 2016

Dissertation directed by: Professor & Dean Emerita, Jennifer Preece,  
College of Information Studies

For several decades, the intensifying trend of researchers to believe that sharing research data is “good” has overshadowed the belief that sharing data is “bad.” However, sharing data is difficult even though an impressive effort has been made to solve data sharing issues within the research community, but relatively little is known about data sharing beyond the research community. This dissertation aims to address this gap by investigating *how data are shared effectively across research and public communities*.

The practices of sharing data with both researchers and non-professionals in two comparative case studies, Encyclopedia of Life and CyberSEES, were examined by triangulating multiple qualitative data sources (i.e., artifacts, documentation, participant observation, and interviews). The two cases represent the creation of biodiversity data, the beginning of the data sharing process in a home repository, and the end of the data sharing process in an aggregator repository. Three research questions are asked in each case:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

The findings reveal the data sharing contexts and processes across research and public communities. Data sharing contexts are reflected by the cross-level data providers and human mediators rooted in different groups, whereas data sharing processes are reflected by the dynamic and sustainable collaborative efforts made by different levels of human mediators with the support of technology mediators.

This dissertation provides theoretical and practical contributions. Its findings refine and develop a new data sharing framework of knowledge infrastructure for different-level data sharing across different communities. Both human and technology infrastructure are made visible in the framework. The findings also provide insight for data sharing practitioners (i.e., data providers, data mediators, data managers, and data contributors) and information system developers and designers to better conduct and support open and sustainable data sharing across research and public communities.

DATA SHARING ACROSS RESEARCH AND PUBLIC COMMUNITIES

by

Yurong He

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2016

Advisory Committee:

Dr. Jennifer Preece, Chair

Dr. Brian Butler

Dr. Kari Kraus

Dr. Andrea Wiggins

Dr. Katherine Stewart (Dean's representative)

© Copyright by  
Yurong He  
2016

## Acknowledgements

Praise and thank my true and living God who is my strength and bestows wisdom and endurance upon me to accomplish these doctoral studies. My sincerest gratitude to many people who provided me with tremendous help during my PhD studies.

First and foremost, I am thankful to Dr. Jennifer Preece for being my incredible advisor over the years. Jenny has patiently guided me through this whole journey, providing me with invaluable advice whenever I needed it. She has always had trust and confidence in me. Her profound academic knowledge and her dedication to the students will have a lasting impact on my life.

Dr. Andrea Wiggins has been a very important mentor to me for the last two years. Andrea opened my eyes to a wider realm of knowledge. She has polished my research abilities, giving me a lot of support in the work of this dissertation. Dr. Brian Butler gave me invaluable advice at different stages of developing this dissertation, advice that was critical to laying the solid foundation of this dissertation in the beginning and raising its quality to a higher level toward the end. Many thanks to other members of my committee, Dr. Kari Kraus and Dr. Katherine Stewart, for their very helpful advice, invaluable feedback, and ongoing support. Likewise, I also thank Dr. David Jacob, a member of my dissertation proposal committee, for the support and advice that helped me develop the proposal into a full thesis.

Dr. Jen Hammock from the Smithsonian Institute is an amazing and longtime collaborator. Jen gave very generously of her time, patience, and resources to support my doctoral studies. Without Jen, my dissertation could not have been accomplished.

I thank Dr. Cyndy Parr and Dr. Derek Hansen for their great support in the early stage of my PhD studies. I also thank Dr. Rob Stevenson for being a wonderful mentor at DataONE and an inspiration for me.

I am greatly fortunate to have supportive research teammates: Carol Boston, Dr. Anne Bowser, and Dr. Dana Rotman; Marina Cardoso, Liz Warrick, and Alina Goldman; Dr. Michele Weber and Dr. Seabird McKeon. Other iSchool doctoral students—Dr. Jinyong Kim, Dr. Beth Bonsignore, Chi Young Oh, Xu Meng, and Lingzi Hong—also offered me their support and encouragement. I thank Fiona Jardine for offering excellent help in editing this dissertation.

I thank my brothers and sisters in the Maryland Chinese Bible Study Group, Chinese Bible Church of College Park, and Plateau True Light Church for their prayers and love. I also thank my roommates, Hsiang-Ling and Yang, for their company.

I am deeply grateful to my parents, Xihua and Yunfen, and my stepparents, Kui and Min, and many other family members. My success would not have been possible without their endless love and support. Last, but definitely not least, I thank my loving husband, Xiao, for being my most intimate spiritual partner. I love you.

# Table of Contents

Acknowledgements.....	ii
Table of Contents.....	iv
LIST OF FIGURES .....	vii
LIST OF TABLES.....	x
1 Introduction .....	1
1.1 Motivation and research questions.....	4
1.2 Conclusion .....	11
2 Background Literature and Theoretical Frameworks.....	13
2.1 What are data?.....	13
2.2 Data sharing mediators .....	14
2.3 Data sharing infrastructures .....	15
2.3.1 Knowledge infrastructures .....	15
2.3.2 Cyberinfrastructure .....	16
2.4 Data sharing challenges .....	23
2.5 Data (not) sharing—scientists and citizen scientists.....	25
2.6 Theoretical and analytical models .....	28
2.6.1 The data life cycle model.....	29
2.6.2 The framework of academic data sharing.....	34
2.6.3 The model of microfunctions of institutional logics.....	46
2.6.4 The model of identification.....	53
2.6.5 Integrate theoretical frameworks .....	55
2.7 Conclusion .....	56
3 Methods.....	60
3.1 Research design .....	60
3.1.1 Making sense of the research design .....	60
3.1.2 Research methods .....	63
3.2 Domain selection .....	65
3.3 Case selection.....	68
3.3.1 Theoretical sampling.....	68
3.4 Selected cases.....	73
3.4.1 Encyclopedia of Life.....	73
3.4.2 The CyberSEES project.....	74

3.5	Data collection and Analysis .....	76
3.5.1	Research procedure .....	76
3.5.2	Data collection .....	77
3.5.3	Data analysis .....	82
3.6	Data validation .....	85
3.7	Conclusion .....	89
4	Case one – Encyclopedia of Life .....	91
4.1	Overview of the findings .....	91
4.2	Answering the first question: who are the data providers?.....	94
4.2.1	Diversity of data providers.....	94
4.3	Answering the second question: who are the data sharing mediators?.....	100
4.3.1	Human mediators .....	100
4.3.2	Technological mediators.....	109
4.4	Answering the third question: what are the data sharing processes?.....	119
4.4.1	The general processes of sharing data .....	120
4.4.2	The data sharing processes vary with each content partner: the influence of human infrastructure.....	132
4.4.3	The number of human mediators and the time for building partnerships...	144
4.5	Conclusion .....	149
5	Case Two – The CyberSEES Project .....	151
5.1	Overview of the findings .....	151
5.2	Answering the first question: who are the data providers?.....	153
5.2.1	Visible data providers in the online environment.....	156
5.2.2	Data creators in the offline environment .....	156
5.3	Answering the second question: who are the data sharing mediators?.....	160
5.3.1	Human mediators .....	161
5.3.2	Technology mediators.....	165
5.4	Answering the third question: what are the data sharing processes?.....	176
5.4.1	Share to home repository .....	176
5.4.2	Share to aggregator repository .....	182
5.5	Conclusion .....	192
6	Discussion.....	194
6.1	Overview of the chapter.....	194
6.2	Summarizing and comparing two cases and their findings regarding the research questions .....	194

6.2.1	Data providers: different types of data providers at different levels .....	204
6.2.2	Human mediators: different structures.....	209
6.2.3	Technology mediators: barriers vs. no barriers.....	212
6.2.4	Data sharing processes.....	216
6.3	Implications for theory: connecting data sharing contexts and processes .....	218
6.3.1	First version integrated framework of data sharing .....	220
6.3.2	Second version of the integrated framework of data sharing .....	226
6.4	Implications for the data sharing practices .....	232
6.4.1	Data sharing practice for research community and for both research and public communities.....	235
6.4.2	Sharing data on the home repository and on the aggregator repository .....	240
6.4.3	Peer reviewing the data created by non-professionals.....	245
6.5	Implications for the design of data sharing.....	250
6.5.1	Preparing ready-to-use data files .....	252
6.5.2	Design for supporting collecting and uploading data simultaneously .....	255
6.5.3	Visible metadata and data quality assessment information .....	257
6.6	Conclusion .....	258
7	Conclusion.....	260
7.1	Contributions .....	263
7.2	Limitations .....	268
7.3	Future papers.....	272
7.4	Future work.....	276
	Appendices.....	280
	Appendix One – Interview slides for the case of EOL .....	280
	Appendix Two – Initial coding schema .....	283
	Appendix Three: The data sharing stories.....	287
	Bibliography .....	305

## LIST OF FIGURES

Figure 1.1. Data download page of the project eBird (Cornell Lab of Ornithology, 2016). .....	7
Figure 1.2. Data download page of the Project BudBurst (Project BudBurst, 2016). .....	8
Figure 1.3 Roundtable model. Cited from Soranno et al., (2015). .....	9
Figure 2.1 Identifying the data sharing mediators in Bietz et al., (2010). .....	21
Figure 2.2 The relationships between the models/frameworks. ....	29
Figure 2.3 The data life cycle cited from Michener and Jones (2012). .....	31
Figure 2.4 The data life cycle cited from Rüegg et al. (2014). .....	32
Figure 2.5 Framework for academic data sharing, cited from Fecher et al. (2015). .....	35
Figure 2.6 The framework of data sharing processes for this doctoral research. ....	45
Figure 2.7 A cross-level model of institutional logics combining macro-micro and micro- macro, cited from Thornton et al. (2012). .....	47
Figure 2.8 The framework of data sharing contexts for this dissertation. ....	53
Figure 2.9 A fuzzy model of identification cited from Ashforth et al., (2008). .....	54
Figure 2.10 The two-way relationship among the models/frameworks. ....	56
Figure 3.1 Research procedure and timeline for this doctoral research. ....	77
Figure 4.1 The number of different types of data providers. The total number of data provider is 329. ....	99
Figure 4.2 EOL organizational chart .....	107
Figure 4.3 The content partner account screenshot. ....	112
Figure 4.4 An example of EOL species page for Giant Panda screenshot (Overview page) (Giant Panda, n.d.). .....	115

Figure 4.5 The TraitBank search result screenshot (Search TraitBank, n.d.).....	118
Figure 4.6 The relationships between data sharing processes, decision making and implementation, and the interrelationships between human actors and their organizations/institutions/communities. ....	134
Figure 5.1 Screenshot of the iNaturalist Biocubes project page (Biocubes, 2015). ....	155
Figure 5.2 Screenshot of the data entry system on iNaturalist. ....	168
Figure 5.3 Screenshot of an organism observation page on iNaturalist (He, 2016). ....	170
Figure 5.4 Screenshot of the basic search and the Filters panel on iNaturalist (Observations, n.d.).....	172
Figure 5.5 Screenshot of the data exporting page on iNaturalist.....	175
Figure 5.6 The Biocubes project participants were using biocube kit to build biocubes. .....	178
Figure 5.7 A Biocubes project participant was taking a photo of a frozen ant.....	181
Figure 5.8 Example of the data quality assessment panel for a research-grade datum on iNaturalist.....	186
Figure 5.9 A Biocubes datum on iNaturalist (drop, 2015, January). ....	190
Figure 5.10 The Biocubes datum (i.e., photograph) is shared from iNaturalist to EOL platform (Ribbed Mussel, n.d.). ....	191
Figure 5.11 The Biocubes datum (i.e., photograph) is shared from iNaturalist to EOL platform, with detailed data source information (drop, 2015, March).....	192
Figure 6.1 Human mediators structure on EOL, an aggregator repository.....	212
Figure 6.2 Human mediator structure on iNaturalist, a home repository. ....	212

Figure 6.3 Adapting the framework of data sharing, model of data life cycle model, the model of microfunctions of institutional logics, and the model of identification into two theoretical frameworks: the framework of data sharing processes and the framework of data sharing contexts.....	219
Figure 6.4 The first version integrated framework of data sharing processes and contexts. ....	220
Figure 6.5 The second version integrated framework of research data sharing. ....	227
Figure 6.6 Data sharing environments and different levels of data providers.....	228
Figure 6.7 Data sharing environments and different level data providers.....	229
Figure 6.8 Data sharing processes facilitated by data sharing mediators. ....	230
Figure 6.9 The relationships between identities and social interactions.....	232
Figure 7.1 Example of the summary of the usages statistics for a content partner.....	277

## LIST OF TABLES

Table 2.1 The differences of using the word “traditional” .....	34
Table 3.1 Two cases selected based on theoretical sampling criteria.....	73
Table 3.2 Summary of multiple data sources and data analysis methods regarding answering the research questions for each case.....	84
Table 4.1 Six types of combinations of relationships among data managers, data contributors, and technicians. ....	144
Table 4.2 The number of human mediators for the different types of content partners.	146
Table 4.3 The numbers of human mediators from EOL and different content partners.	148
Table 4.4 The time the 18 content partners took to build the partnerships with EOL....	148
Table 5.1 The data creators' organizational identities.....	157
Table 5.2 The Biocube project organizers’ organizational identities .....	163
Table 5.3 Biocubes data collected in the Florida workshop by the data providers. The total number of data providers are not available because there are overlapping data providers for different days.....	188
Table 6.1 The differences between the two cases from the perspective of data sharing contexts and data sharing processes.....	195
Table 6.2 Summary of the major differences between the knowledge infrastructures of the two cases. ....	198
Table 6.3 The comparison of findings between case one (EOL) and case two (CyberSEES).....	203
Table 6.4 Different level of data providers in the case of EOL and the case of the CyberSEES project. ....	204

Table 6.5 The human mediators in EOL and Biocubes.....	209
Table 6.6 The technology mediators in EOL and CyberSEES.....	213
Table 6.7 The general steps included in the data sharing processes for EOL and CyberSEES. ....	216

## 1 Introduction

For several decades, the intensifying trend of researchers to believe that sharing research data is “good” has overshadowed the belief that sharing data is “bad.” A scientific culture of “extreme openness” is developing causing all information of scientific value—from raw data and computer code to questions, ideas, folk knowledge, and speculation—to become available on public networks (Nielsen, 2012; Edwards et al., 2013). In March 2015, National Science Foundation’s (NSF) released its two-year public access plan, *Today’s Data, Tomorrow’s Discoveries*, to “expand public access to the results of its funded research” (NSF, 2015).

Borgman (2012) has identified four benefits for sharing data, including “to make results of publicly funded research available to the public,” “to reproduce or to verify research,” “to enable others to ask new questions of extant data,” and “to advance the state of research and innovation” (p. 1059). However, sharing data has long been known to be difficult (Kowalczyk & Shankar, 2011). A new data sharing culture together with funding agencies’ explicit requests to share data and provide public access contribute to the significant crossroads that the research community finds themselves now at (Fenichel & Skelly, 2015).

Public access to data means not just sharing data with academic researchers, but also with data users outside the academy (i.e., outside the research community) (Fenichel & Skelly, 2015). Previous research describes the impressive effort made to solve data sharing issues within the research community by developing cyberinfrastructure, professional

repositories, metadata, and various tools that assist in sharing, aggregating, and integrating research data (Kowalczyk & Shankar, 2011). However, relatively little is known about data sharing beyond this community (Soranno, Cheruvellil, Elliott, & Montgomery, 2015). Questions, such as when and how to share data and what value there is in sharing data with the public, remain unanswered. This dissertation aims to address this gap in our knowledge by investigating *how data are shared effectively across research and public communities*.

Based on the definition of data sharing by Soranno et al. (2015), this dissertation defines *data sharing* as sharing research data or data that has the potential to become research data in any publicly accessible repository. *Research community* is defined as a group of researchers or organizations of any size and can be discussed with a general or specific meaning, or at macro, meso, and micro levels. The primary goal of using data by research community members is to solve research problems. Similarly, *public community* refers to a group of members or organizations of any size from the general public and can be discussed with a general or specific meaning, or at macro, meso, and micro levels. Most members from the public community are non-professionals. They could use data for any purpose. *Effectively* means that data sharing actually occurred and the data are successfully published on one or more public accessible repositories.

Data sharing has greatly advanced a few data-intensive science disciplines already, such as genomics (Kaye, Heeney, Hawkins, De Vries, & Boddington, 2009), meteorology (Hayes, 2012), astronomy (Ivezic, 2012), health (Piwowar, Becich, Bilofsky, & Crowley,

2008), neuroscience (Van Horn, 2008) and benefitted society in general (Soranno et al., 2015). Research data in those, and many other fields, are not only contributed by researchers, but also by non-professionals. In recent years, increased attention has been focused on the paradigms in which non-professionals participate as both data contributors and users, collecting, visualizing, analyzing, and learning from research data (Dickinson et al., 2012). In these ways, non-professionals have played an increasingly important role in the progress of science.

There has been a large body of data sharing research in different scientific domains, including traditional as well as relatively new interdisciplinary arenas, such as Computer-Supportive Cooperative Work (CSCW) and Human-Computer Interaction (HCI). However, most prior research has focused on data sharing within the research community. The question of how to share data with the general public as well as researchers has not yet been well answered.

To make data sharing possible, data creators must first be willing to share their data. With social and technological support, creators become providers by making their data available to the users. Social and technological support components of data sharing can be collectively referred to as *data sharing mediators*. As an entry point to investigate the practices of making data sharing possible for both research and public use, this dissertation considers the work of mediators that connect data creators and data users.

Social support refers to any efforts made by human actors in specific social situations to achieve the work of data sharing. These efforts include making the data shared from different providers become meaningful to users. Technological support includes any information and communication technologies (ICTs) that are used or specifically designed for enabling and facilitating data sharing, such as data sharing tools and repositories. These data sharing mediators are responsible for building and maintaining the relational, ecological, and sustainable knowledge infrastructure (Starr, 1999; Ribes & Lee, 2010) that network “people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds” (Edwards, 2010, p. 17).

This dissertation contributes to the promotion of data sharing culture, the understanding of data sharing practices, and refining a data sharing framework, as well as provides insights about practice and design for sharing data across research and public communities.

## 1.1 Motivation and research questions

The call for improving data sharing has existed since the 1980s (Borgman, 2012). As the age of data deluge, or “big data,” has arrived, an increasing number of institutions, organizations, and communities across disciplines and domains have realized the value of big data to address major scientific and social issues (e.g., climate change, biodiversity triage, and health care) and therefore the urgency of improving data sharing (Borgman, 2012; Edwards. et al., 2013). Funding agencies such as the NSF and the National

Institutes of Health (NIH), research institutions, and many peer-reviewed journals have taken actions to put this call in their data management policies and requirements (Borgman, 2012). These laws and policies have been considered as the most motivating external social influences on data sharing (Zimmerman, 2003). However, the practice of science still comes into conflict with its often-mentioned values of openness and shared data (Kowalczyk & Shankar, 2011).

There has been extensive research on examining data sharing conditions and challenges. Except for a few science domains mentioned above that have developed data sharing cultures and norms with relative success, many other domains have not been ready to shift toward these norms of data sharing with other researchers (Soranno et al., 2015; Borgman, 2015), let alone the general public. The obstacles of data sharing range from technological, institutional, legal, financial, to cultural and behavioral (Arzberger, et al., 2004; Kowalczyk & Shankar, 2011). Although these challenges are not easy to address in the short time, some researchers are making progress by developing various tools and strategies, such as cyberinfrastructure, data repositories, data sharing tools, metadata standards, data publication, and attribution and professional credit systems (Parr & Cummings, 2005; Goring et al., 2014; Soranno et al., 2015).

While these obstacles are challenging enough, Soranno et al. (2015) indicate that a larger hurdle of data sharing is that “there appears to be no strong ethical impetus for sharing data within the current culture, behaviors, and practices of scientists” (p. 70). This challenge has not been addressed as well as others. Many researchers do not think it is

their obligation to share their data with others, especially beyond their inner network or community (Soranno et al., 2015). For some researchers, especially those who have worked long and hard to collect it, their data are like their babies. They are reluctant to give their babies to others, especially to those who have little direct relationship (e.g., members from the public) with them and their babies. When there are policies that require them to share data, they do not perceive sharing as a “cheerful” task.

However, data sharing in citizen science projects shows a different and much more promising picture. Citizen science is a special type of research and practice that involves members of the public, usually non-professionals, collecting and/or analyzing scientific data (Bonney et al., 2009). These non-professionals, “citizen scientists,” do not have to have professional research training before they participate. Most citizen science projects share their data with not only scientists, but also the general public. For example, the project eBird allows the public to freely access and download their database (Figure 1.1). This is also a way of offering the data as feedback and reward to the contributing citizen scientists. Project BudBurst also makes their metadata freely available by allowing the data to be downloaded and used for noncommercial purposes (Figure 1.2).

In these citizen science project examples, data sharing has not been prevented by any challenges mentioned above. By focusing their attentions on the social consciousness and democratization of science, as well as the inclusion of non-professionals from the public (Soranno et al., 2015), these projects take data sharing with both the research and

the public community as integral to their mission: they view sharing data with the public as their ethical obligation.

**eBird**

Home About Submit Observations Explore Data My eBird Help

Sign In or Register Language

## View and Explore Data

**Explore a Region**  
Recent sightings, checklists, birding activity, best hotspots, and top birders for a county, state, province, or country.

**Explore Hotspots**  
Discover the best places for birding nearby or around the world.

**Species Maps**  
Explore interactive range maps by species or subspecies — zoom in for details

**Bar Charts**  
Find out what birds to expect throughout the year in a region or location

**Line Graphs**  
Explore different metrics of species occurrence in a region or location

**Submission Map**  
Watch in real time as people submit their sightings from across the globe

**Arrivals and Departures**  
Arrivals and departures for a country, state/province, county, or hotspot

**All-Time First/Last Records**  
All-time records for species arrival and departure in a region

**High Counts**  
Species high counts for a region

**Summary Tables - Your Data**  
Your observations summarized by week, month, or year

**Download Data**  
Download eBird data to use in your own projects.

**Your Totals**  
Track your totals and compare with other eBirders.

**Yard Totals**  
How many species and checklists have you submitted for your yard?

**Patch Totals**  
How many have you submitted for your favorite birding patches?

**Top 100**  
Compare with the top eBirders in your region.

**Species You Need**  
Tools to find species you haven't seen yet.

**Target Species** NEW  
Prioritized list of county, state, or life birds that you can expect to find in a region

**Alerts**  
Reports and email alerts for rarities and species you haven't seen

Connect with eBird — Subscribe to our Email Newsletter

Enter your email address...

The latest news about eBird, birding, ornithology, and conservation delivered to your inbox.

© Cornell Lab of Ornithology | Contact | FAQ

Figure 1.1. Data download page of the project eBird (Cornell Lab of Ornithology, 2016).

Project BudBurst Data

The 2016 campaign is underway! The latest observations can be viewed in the [live map](#) shown on the [Results](#) page. Also you can view [live maps](#) of each plant species by visiting our [View All Plants](#) page and selecting a plant from the list.

**Project BudBurst Data Citation and Community Attribution**

Project BudBurst data is freely available for anyone to download and use for noncommercial use. The data is provided by thousands of observers from across the country. Please cite your use of the data and recognize our observers with the following citation and [community attribution](#):

Project BudBurst. 2016. Project BudBurst: An online database of plant phenological observations. Project BudBurst, Boulder, Colorado. Available: <http://www.budburst.org>; Community Attribution: [http://www.budburst.org/results\\_attribution](http://www.budburst.org/results_attribution); Accessed: April 15, 2016.

**Project BudBurst Data Downloads**

Year	Datasets	Metadata	Reports
2015	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2015 Metadata (.xlsx)</a>	
2014	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2014 Metadata (.xlsx)</a>	
2013	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2013 Metadata (.xlsx)</a>	
2012	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2012 Metadata (.xlsx)</a>	
2011	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2011 Metadata (.xlsx)</a>	<a href="#">2007-2011 Summary Report (PDF)</a>
2010	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2010 Metadata (.xlsx)</a>	
2009	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2009 Metadata (.xlsx)</a>	
2008	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2008 Metadata (.xlsx)</a>	<a href="#">2008 Summary Report (PDF)</a>
2007	<a href="#">.xls</a>   <a href="#">.csv</a>   <a href="#">.txt</a>	<a href="#">2007 Metadata (.xlsx)</a>	<a href="#">2007 Analysis Report (PDF)</a>

Depending on your Operating System and Web Browser, you may need to right-click on the links and "Save As..." to download the file to your computer.

Figure 1.2. Data download page of the Project BudBurst (Project BudBurst, 2016).

The promising situation of public sponsorship and participation in research is accompanied by the growing expectation and even requirement of funding agencies for researchers to share their data (Soranno et al., 2015). This combination pushes researchers, especially those who seek to broaden public participation in science, towards viewing data sharing as an ethical obligation, albeit often limited to researchers and not citizen scientists or the public at large (Soranno et al., 2015).

Based on the lessons learned from citizen science, Soranno et al. (2015) propose a roundtable model in which not only scientists, but also the public, policymakers, and

stakeholders are invited to sit around a metaphorical, sometimes actual, table to become more involved in science (Figure 1.3). This model has become more common in recent years. It shows that publicly sharing data and making it accessible for everyone benefit different aspects of the publicly funded science system (e.g., data sharing practices, public awareness of science, and policymaking). By sitting around the same table, these different communities are connected and no longer isolated. Therefore, Soranno et al. (2015) call for “a shift toward the ethical value of promoting inclusivity within and beyond science,” and emphasize that “an essential element of a truly inclusionary and democratic approach to science is to share data through publicly accessible data sets” (Soranno et al., 2015, p. 1).

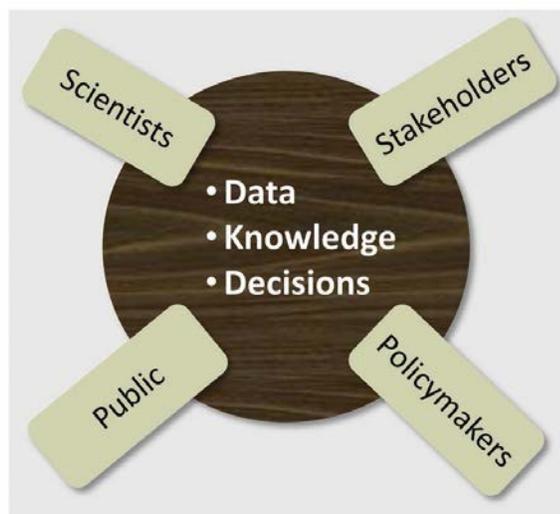


Figure 1.3 Roundtable model. Cited from Soranno et al., (2015).

This dissertation responds to Soranno et al.’s (2015) call by investigating how data can be effectively shared through publicly accessible data sets. Soranno et al. (2015) strongly argue that sharing data with the general public as well as researchers plays a critical role in publicly funded science. However, when it comes to online environments and sharing

data with the general public, it is unclear what constitutes effective practices. Finding answers to this was the major motivation behind conducting this research. This dissertation extends the scope of data sharing from the research community to the public community. Stakeholder and policymaker communities are not considered in the current study, but should be included in future research. Therefore, the overarching research question of this study is:

***How data are shared effectively across research and public communities?***

As mentioned earlier, any data sharing needs mediators to connect creators with users, allowing data to flow from one to the other, and generating meaning and utility for users (Borgman, 2015). When the data creators are willing to share their data, they gain a new identity: *data providers*. The mediators' role is to create and manage reliable online ecosystems to support data sharing from providers to users. Data providers and mediators are central actors in data sharing practices in order to share data effectively across research and public communities in online environments. As data sharing practices can be reflected by specific contexts and processes, this dissertation breaks down the overarching question into three research questions:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

The first two questions were asked in order to reveal data sharing contexts. To answer all three questions, two real-world cases in which data are effectively shared across research and public communities were analyzed using a comparative case study method. The cases were carefully selected by following guidance from pre-selected theoretical and analytical frameworks.

## 1.2 Conclusion

This chapter introduced the motivation behind conducting the research in this dissertation, the overarching research question, and the three specific research questions that stem from the overarching question. This chapter also introduced a set of concepts (i.e., data providers, data sharing mediators, data sharing contexts, data sharing processes, data sharing, research community, and public community) which will be discussed in more detail in the following chapters. This dissertation organizes its contents as follows:

**Chapter 2 – Background literature and theoretical framework.** This chapter first introduces the literature including cyberinfrastructure development and data sharing challenges in the research community. It presents a set of theoretical frameworks that guided the case selection and initial qualitative data analysis.

**Chapter 3 – Methods.** This chapter introduces the rationale of the research design and presents details about the comparative case study method of research. Data collection and analysis are also described in this chapter.

**Chapter 4 & Chapter 5 – Case one (Encyclopedia of Life) and Case two (The CyberSEES project).** The findings of the two cases (i.e., responding to the three research questions) will be presented in these two chapters respectively.

**Chapter 6 – Discussion.** This chapter will present the comparison between the characteristics of the two cases and the findings of the two case studies. Based on these findings, a new integrated theoretical framework of research data sharing was developed. To conclude, this chapter discussed the valuable implications for data sharing practices and designs.

**Chapter 7 – Conclusion.** This chapter summarizes the key findings, discusses the limitations of this study, and future directions.

## 2 Background Literature and Theoretical Frameworks

### 2.1 What are data?

Data can be interpreted differently by different individuals and groups in different contexts (Borgman, 2015). The differences start with defining what data are. The concept of data is usually discussed with that of information and knowledge since they are three fundamental building blocks in the field of information science (Zins, 2007). Zins (2007) examines the different definitions of data, information, and knowledge suggested by 57 leading information science scholars from 16 countries. He points out that data can be “used in the plural or as a singular word meaning a set or collection of facts” (p. 481), and is defined most frequently as symbols, sensory stimuli, disconnected facts, and observations (i.e., raw data) (Zins, 2007). The basic relationship among these three basic blocks is that information is a useable form of processed data providing answers to the “who,” “what,” “where,” “when,” and “how many” questions; and that knowledge is the application of data and information to answer “how to” questions (Ackoff, 1989). This dissertation uses “data” as a plural meaning a collection of facts and the singular, “datum,” to mean a single fact.

Data is a broad concept that is difficult to define more specifically, especially considering the many different contexts in which it is used. Following Borgman (2015), this dissertation narrows the concept of data in the context of scholarly communication: when the concept of data is discussed, it refers to either research data or data that could be or have potential to become research data. Furthermore, in an operational as well as a general research context, the most useful definition of data is descriptive of the

categorization of data in practical ways, such as grouping by origin, value, or other factors (Borgman, 2015). One widely accepted categorization of data was developed by the US National Science Board to reflect types of data in the sciences, social sciences, and technology; its four general categories are observational (e.g., data are directly observed and collected by human or machine sensors), computational (e.g., data are created via computer modeling, simulations, etc.), experimental (e.g., data are created in scientific experiments), and records (e.g., artifacts, documentations) (National Science Board, 2005; Borgman, 2015).

## 2.2 Data sharing mediators

The term “mediator of data sharing” is decades old and refers to an approach, virtual database, and system to integrate data from diverse databases, thereby connecting data sources and the application (i.e. computer program) using them (Wiederhold, 1992).

However, as the importance of human infrastructure to enable data releasing, sharing, and reusing has garnered greater recognition in recent years, this term has also been used to refer to the people who connect data creators and data users (Borgman, 2015). The work of human mediators, such as those curating and managing data, developing and maintaining sharing standards and technologies is crucial in making data sharing possible; nevertheless, human mediators are usually invisible and overlooked (Kervin, Cook, & Michener, 2014; Borgman, 2015). These two concepts of mediator are both accurate, but technological components (e.g., information systems) have consistently gained most attention and investment compared to human and social components in terms

of efforts made to develop and improve science and other knowledge infrastructures (Edwards et al., 2013).

In recent years, demands to redress this imbalance are increasing. For example, in some fields, researchers agree that solving the sociological challenges of data sharing and integration is even more important than solving technical challenges (Parr, Guralnick, Cellinese, & Page, 2012). As the advancement of information and communication technology enables increasingly complex and broader collaboration and cooperation among people, groups, and organizations than decades ago, it is impossible to build a successful knowledge infrastructure without carefully considering the influence of individuals, society, culture, organizations, and institutions (Edwards et al., 2013). The human mediators' practices surrounding the development and maintenance of various knowledge infrastructures are critical for better connecting data creators and data users. The concept of a mediator of data sharing is therefore a sociotechnical one.

## 2.3 Data sharing infrastructures

### 2.3.1 Knowledge infrastructures

Knowledge infrastructures are defined as “robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds.” (Edwards, 2010, p. 17). Frequently, infrastructure related to data sharing is envisioned as an information system that includes hardware, software, data formats and protocols (Geschwind, 2001; Kowalczyk & Shankar, 2011). However, the definition of knowledge infrastructure employs a broader theoretical understanding of infrastructure.

Here infrastructure does not merely focus on “what,” but places more emphasis on sustained relationships (Ribes & Lee, 2010). As Star (1999) argued, “infrastructure is both relational and ecological – it means different things to different groups and it is part of the balance of action, tools, and the built environment, inseparable from them” (p. 377).

### 2.3.2 Cyberinfrastructure

Cyberinfrastructure is a typical knowledge infrastructure consisting of networked computational tools, resources, and collaborative efforts, focusing on supporting scientific research activities (e.g., interdisciplinary collaboration, data sharing, dissemination of findings) (Ribes & Lee, 2010; Atkins, 2003). Cyberinfrastructure’s major mission is to revolutionize science, with communities of researchers as its target (Atkins, 2003; Ribes & Lee, 2010). These infrastructures provide scientists the features they need to answer their research questions (Bietz, Baumer, & Lee, 2010). However, no matter what features could be provided, the answerability of research questions largely depends on whether data are successfully collected and analyzed. Therefore, it is important that one set of features of cyberinfrastructures be data centered, so as to support its collection, analysis, and/or sharing.

“How to facilitate data sharing” is one of the three rapidly growing investigative areas in cyberinfrastructure studies and the only directly related to data (Ribes & Lee, 2010).

When developing a cyberinfrastructure, an ultimate goal and one of the core working practices is building data repositories to facilitate data sharing (Lee, Dourish, & Mark,

2006; Bietz et al., 2010). Therefore, this dissertation focuses on data sharing supported by this infrastructure. More specifically, it addresses the collaborative efforts that support data sharing within knowledge infrastructures.

However, as mentioned in Chapter 1, data sharing should be not only by and for researchers, but also by and for members of the general public. The knowledge infrastructures studied in this dissertation value supporting sharing by and for members of the public as at least as important as by and for researchers, if not more important. As a result, these kind of knowledge infrastructures might not be considered as traditional cyberinfrastructures mainly developed for supporting research work. Nevertheless, studies within cyberinfrastructure studies related to building and managing infrastructure and “how to facilitate data sharing” are a valuable resource given the limited number of studies about types of knowledge infrastructures other than scientific ones (i.e., cyberinfrastructure) that this research can build on.

### 2.3.2.1 Human Infrastructure

It is easy to associate cyberinfrastructure with computing technologies (i.e., technology infrastructure), but a less visible part—human infrastructure—is essential for enabling a whole infrastructure to emerge and function (Lee et al., 2006; Bietz et al., 2010): there is always a human actor’s decision behind every creation of cyberinfrastructure. The phrase “human infrastructure of cyberinfrastructure” was first coined by Fran Berman, the former director of the San Diego Supercomputer Center (Lee, Bietz, & Thayer, 2010):

“The cyberinfrastructure’s human infrastructure is a synergistic collaboration of hundreds of researchers, programmers, software developers, tool builders, and others who understand the difficulties of developing applications and software for a complex, distributed, and dynamic environment. These people are able to work together to develop the software infrastructure, tools, and applications of the cyberinfrastructure. They provide the critical human network required to prototype, integrate, harden, and nurture ideas from concept to maturity. (Berman, 2001)” (Lee et al., 2006, p. 483-484)

Lee et al. (2006) subsequently modified the definition to “the arrangements of organizations and actors that must be brought into alignment in order for work to be accomplished” (p. 484). Instead of having a uniform organizational form (e.g., organizations, networks, or teams), participation in human infrastructure may take more than one or even all of these forms at the same time, indicating human infrastructure’s complex and heterogeneous collaborative structures (Lee et al., 2006; Bietz et al., 2010).

Bietz et al. (2010) expanded the concept of the human infrastructure of cyberinfrastructure by drawing attention away from diverse human collaborative structures and towards the social-technical collaborations that are vital for success in developing cyberinfrastructure. They studied a case of cyberinfrastructure development for metagenomics research by investigating the work of creating the infrastructure, focusing on “the process of purposeful human action” (Bietz et al., 2010, p. 250). In their study, a large-scale, multi-year project named Community Cyberinfrastructure for

Advanced Marine Microbial Ecology Research and Analysis (CAMERA) was chosen as the study subject. The CAMERA project aimed to provide cyberinfrastructure tools and resources and bioinformatics expertise to the metagenomics community (Bietz et al., 2010).

Building a community repository that could be populated with data from scientists and other databases was the key activity when creating infrastructure in the CAMERA project (Bietz et al., 2010). Bietz et al. (2010) discovered that the processes of building the community repository not only included technical level activities (e.g., building scripts), but also required establishing and managing a complex set of social-technical relationships. Furthermore, they found that technical and social-technical level activities need project members (e.g., database developers) to align and leverage the relationships within and across multiple organizational structures. Aligning relationships refers to the work of enacting a relationship between different entities, such that the relationship can produce and function within the nascent cyberinfrastructure, whereas leveraging relationships refers to using an existing relationship with a person, organization, or artifact to build or strengthen another relationship with other people, organizations, or artifacts (Bietz et al., 2010). Aligning and leveraging are considered two subprocesses of the “synergizing”, the key mechanism that connects social and technological aspects of infrastructures in the entire cyberinfrastructure (Bietz et al., 2010).

In the case of the CAMERA project, Bietz et al., (2010) found that synergizing has a special and close relationship with embeddedness, one of the properties of infrastructure:

“synergizing both depends upon and produces embeddedness” (p. 271). Specifically, they found that the work of database developers in the CAMERA project drew on and extended the complex, multi-dimensional network of social and technical relationships within which the database is always and already situated.

In terms of the process of building the community repository, Bietz et al. (2010) focused on investigating the work of CAMERA database developers. They also mentioned other project staff, such as a senior administrator who helped forge the relationships with the scientists who had the data. These scientists agreed to import their data to the community repository of CAMERA: some of the scientists were funded by the same source as the CAMERA project and their grants required the scientists to share their data publicly through the CAMERA community repository. Working with the domain experts (i.e., scientists) allowed the database developers to ascertain the database schema and mechanisms to ensure that the data were useful for data users (i.e., scientists in the same research community) in answering their research questions. Bietz et al. (2010) categorized the efforts made by the human actors when developing the community repository into three groups: importing data, metadata, and landscaping data.

According to the concept of a data sharing mediator introduced above, in Bietz et al. (2010)’s study, the human actors (i.e., the developers, the administrator, the domain experts, and the established repository) can all be considered human mediators. The technology actor is the technology mediator. Respectively, these are the core parts of human infrastructure and technology infrastructure that are responsible for facilitating

data sharing via cyberinfrastructure. Figure 2.1 shows the data sharing mediators identified in Bietz et al. (2010). These different elements are involved—connected even—in the processes of synergizing within cyberinfrastructure development.

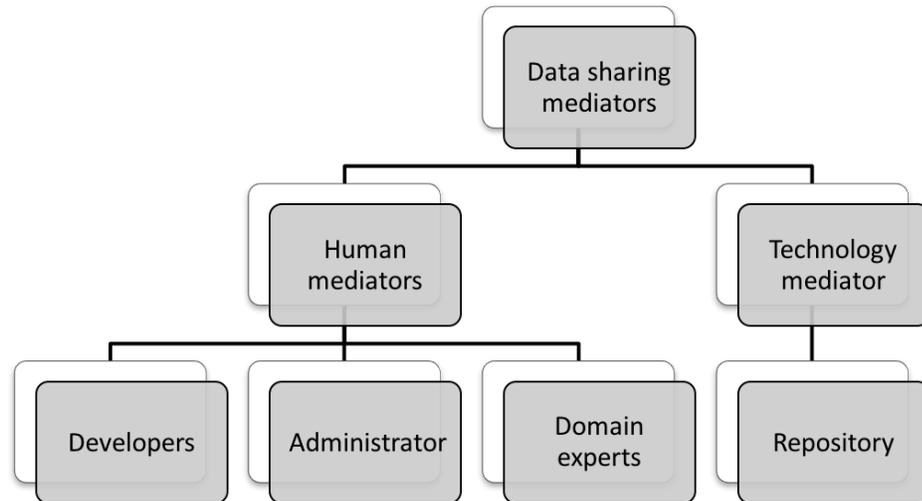


Figure 2.1 Identifying the data sharing mediators in Bietz et al., (2010).

It is worth drawing attention to Bietz et al. (2010)’s use of the term “developer” to refer to anyone who purposely make efforts to create cyberinfrastructure, instead of limiting the definition of developer to programmers who write code to develop software and hardware. Having a general term for the human workers who build the infrastructure is necessary: “developer” is an easy-to-understand and convenient term. However, while Figure 2.1 shows the different identities within which human mediators could be identified, the specific identities of the human actors who create cyberinfrastructure and where these identities originate are still not clear in Bietz et al. (2010).

This dissertation argues that without recognizing the human workers' specific identities, the process of building an infrastructure cannot be fully understood. The term "developer" is insufficient. As in many infrastructure instances, the infrastructure studied in this dissertation contains many different human workers coming from different organizations and institutions and given different organizational identities. They work together, take different responsibilities, and play different roles when creating an infrastructure. Their identities could have important influences on the process of creating the cyberinfrastructure.

In addition, using the term "developer" could be somewhat misleading. On one hand, Bietz et al. (2010)'s study only referred to the human workers within the cyberinfrastructure development team. The data providers (e.g., scientists who provide data to the community repository) were not considered developers because they did not belong to the cyberinfrastructure development team, only provided data, and did not directly work with the programmers on developing the cyberinfrastructure (e.g., writing code to link data to the repository). But in other infrastructure cases, data providers not only provide data but also work with programmers from the cyberinfrastructure team on its development. Therefore, data providers should also be considered as "developers."

On the other hand, creating an infrastructure is not a one-time event. The longevity of the infrastructure must be considered by the development team (Steinhardt, 2016) since infrastructures are supposed to embrace sustainability. Bietz et al. (2010)'s study focuses on the development stage of the cyberinfrastructure. But the day-to-day work for most

people working on an infrastructure is maintenance of the infrastructure, with these people being very likely those who create the infrastructure. Therefore, the “developers” could also be the “maintainers.”

Therefore, this dissertation adopts the term “human mediator” to refer anyone who is directly and intentionally involved in developing and maintaining a knowledge infrastructure that supports data sharing. This term is not limited to the human workers who belong to the infrastructure development team. This dissertation identifies the specific identities of these human workers in order to better understand the process of building and maintaining an infrastructure.

## 2.4 Data sharing challenges

Sharing research data is a challenging task, whether for the data providers or human mediators. There are certain challenges, both technical and social, that need to be addressed by the human workers in any examination of research data access and sharing regimes (Arzberger et al., 2004).

The technological challenges of data sharing are predominantly rooted in developing, applying, and adopting information and communication technologies, such as cyberinfrastructure, repository, metadata, and various tools that enable broad access to and optimal exploitation of research data (Arzberger et al., 2004; Kowalczyk & Shankar, 2011). These type of challenges are usually the result of the nature of data. For example, for ecological informatics, the three major technological challenges are data dispersion,

heterogeneity, and provenance (Reichman, Jones, & Schildhauer, 2011). Although large amounts of scientific data have been digitized and stored somewhere on the Internet by tens of thousands of researchers, these data “remain[] scattered, poorly documented, and in formats that impede discovery and integration” (Parr et al., 2012, p.94).

One explanation for the data heterogeneity that makes it difficult to compare and integrate different data sets is the variety of experimental methods researchers use to collect data across a wide range of topics. The challenge of capturing information about data provenance (i.e., origin and history), especially after data have been subjected to complex and multistep processes during collection and/or analysis, may cause concern about data quality. These challenges need to be addressed by new and powerful technological solutions (Reichman et al., 2011).

Difficulties in establishing and maintaining collaboration and cooperation among human actors lay at the core of the social challenges faced by data sharing, reminiscent of work patterns and content of the human infrastructure of cyberinfrastructure discussed in Section 2.3. One response to these challenges is the effort of domain experts in building communities of cooperation and promoting a culture of community within them (Parr et al., 2012). In these communities, the research scope of members might go beyond that of the human infrastructure of cyberinfrastructure and thus researchers with broader interests are also included.

These social challenges can be understood at two levels: macro and micro. Macro level social challenges include cultural, institutional, organizational, law, policy, ethical, financial and budgetary, and managerial and require addressing when data sharing facilitators develop and maintain data sharing knowledge infrastructures (Arzberger et al., 2004; Kowalczyk & Shankar, 2011). Micro level social challenges include motivation of individual actors as data providers to share data, and as data users to use data (Arzberger et al., 2004). These two level of challenges are not exclusive, but intertwine and influence one other. For example, the variety of institutional models and tailored data management approaches that are most effective in meeting the needs of researchers, and laws, policies, and agreements directly affect data access and sharing practices by individual actors (Arzberger et al., 2004).

## 2.5 Data (not) sharing—scientists and citizen scientists

This dissertation not only examines data sharing for the research community, but studies data sharing for the public community. Data creators that are willing to share their data and become data providers. These individual data creators/providers can be either researchers or non-professionals (Soranno et al., 2015). Non-professionals include members of the general public who do not necessarily have professional research training or work in a research position. As data creators share their data, regardless of its form, data sharing mediators ensure that data are adequately represented and documented within the knowledge infrastructure (Borgman, 2015). Literature about why data creators choose to share—or not share—their data provides more insight into the relationship between the macro and micro level social challenges of data sharing.

Researchers are undoubtedly the primary and dominant data creator. However, because citizen science has flourished thanks to recent advancements in information and communication technologies, the number of non-professionals who contribute data significantly increases every day and therefore play an increasingly important role in conducting research as compared to decades ago, especially in data-intensive subject disciplines, such as biology and astronomy. Citizen science involves the public in research and builds partnerships between researchers and the public (Bonney et al., 2009; Louv & Fitzpatrick, 2012; Miller-Rushing, Primack, & Bonney, 2012). Citizen science provides opportunities in which researchers and non-professionals meet up and collaboratively collect and process data in an offline environment (Silvertown, 2009). The development of the Internet and mobile computing technologies has also enabled various forms of virtual collaboration between scientists and non-professionals, turning citizen science into technology-supported citizen science (Wiggins, 2012).

According to data sharing literature, sharing challenges mainly originate with researchers. Although sharing data has been heavily promoted for years, its “dirty little secret” is that little sharing may actually be taking place (Borgman, 2011). Even though most present-day researchers would say yes to the question of whether they are willing to share their data, willingness does not equal action (Borgman, 2015). Hampton et al. (2013) conducted a survey of ecological papers from randomly chosen NSF Division of Environmental Biology awards between 2005–2009 to determine how much data were publicly available. They found that ecological data are not typically made publicly

available. Within papers that produced data, less than 50% shared some or all of the data, with sharing mostly taking place through GenBank or TreeBASE, databases designed to encourage access within the research community. Only 8% of papers shared their non-genetic data with the public, indicating that while sharing data within the research community is hard, sharing it with the public is even harder.

Tenopir et al. (2011) gathered responses from 1329 scientists across multiple disciplines regarding their current data sharing practices as well as barriers and enablers to sharing. Their results show that scientists usually do not make their data publicly available in online environments; the two leading reasons for this being “insufficient time” and “lack of funding”. They found that other barriers include “having no place to put the data,” “lack of standards,” “sponsor does not require,” and their data “should not be available” to others. Although these major barriers are difficult to solve, systems that make data sharing quick and easy without additional cost may help (Tenopir et al., 2011).

Soranno et al., (2015) summarized the obstacles that might prevent researchers from sharing data, including insufficient rewards and incentives, concerns about the future study being “scooped” (Reichman et al., 2011; Wolkovich, Regetz, & O'Connor, 2012; Goring et al., 2014), technological challenges of data sharing (Reichman et al. 2011), and “no strong ethical impetus for sharing data within the current culture, behaviors, and practices of scientists” (p. 70).

Non-professional data contributors within citizen science (i.e., citizen scientists) are unlikely to have the reasons to not share data as researchers do. The essence of citizen science is data sharing, without which researchers would not receive the data contributed by citizen scientists. Sharing this data is the most typical form of collaboration between citizen scientists and researchers (Wiggins & Crowston, 2011). If a non-professional data creator does not want to share data, s/he would be unlikely to participate in a citizen science project from the outset.

However, this willingness to share data does not mean that citizen scientists have no concerns about it. A major concern of theirs is privacy (Bowser et al., 2014). For example, since it is commonplace in citizen science projects to collect and share data through citizen scientists' sensor-rich smartphones, one potential cause for concern is the GPS and other personal information that might be attached to that data (Cohen, 2008; Kim, Mankoff, & Paulos, 2013). Luckily, this particular concern is easy to solve by allowing citizen scientists to choose how much they want to share (Cohen, 2008).

## 2.6 Theoretical and analytical models

This dissertation focuses on real-world cases of data sharing to study how data can be shared with the public and how data sharing mediators accomplish that work. Theoretical and analytic models are needed to guide the selection of cases as well as the initial data analysis to understand the contexts within which the data providers and data sharing mediators operate to reach a better understanding of data sharing processes. Four models

are introduced here: (1) data life cycle; (2) academic data sharing; (3) microfoundations of institutional logics; and (4) organizational identification.

The data life cycle model presents the stages of life of research data. The academic data sharing framework depicts the major components of human and data infrastructures that represent the phenomenon of academic data sharing. The microfoundations of institutional logics model describes the interrelationships between individual and organizational social actors when reflecting and influencing each other’s social situations. Lastly, the model of identification, which represents institutional logics within which the social actors are embedded, illustrates organizational identity at different levels. The first two theoretical underpinnings are adopted for understanding data sharing processes, and the third and fourth are adopted for understanding data sharing contexts. Figure 2.2 shows the relationships among these four models/frameworks.

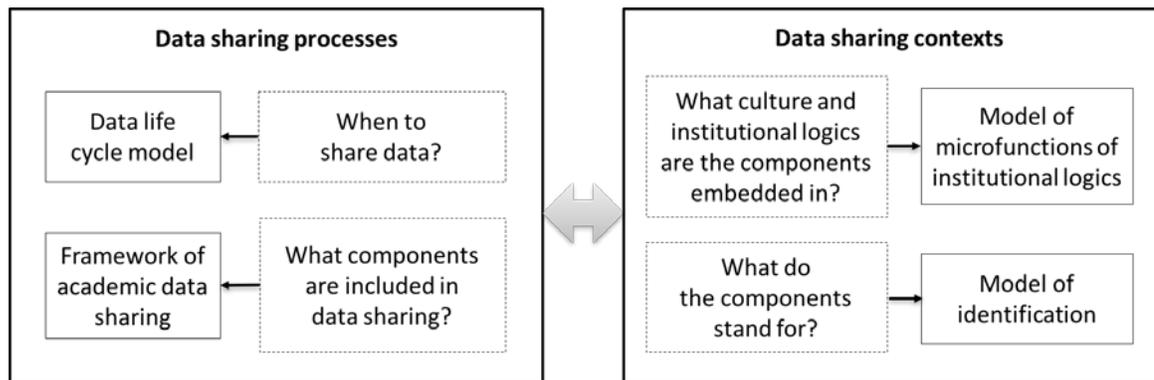


Figure 2.2 The relationships between the models/frameworks.

### 2.6.1 The data life cycle model

Since the 1990s, the data life cycle model has been utilized and improved upon to support digital data preservation and curation practices (Corti et al., 2014). Quoting Michener

and Jones (2012), “The data life cycle encompasses all facets of data generation to knowledge creation” (p. 85). This model can help identify at what stage of the life cycle the data are shared.

Before information technologies were widely adopted in the academic world (i.e., the 1990s), the life of most data would end at some point after the data is published (Michener et al., 1997). As information technologies were developed during the 2000s, a new culture of data sharing appeared and gained popularity in the academic world. Data sharing ceased to be limited to peer-review articles, since numerous data sharing tools and thousands of digital data repositories significantly extended the ability of researchers to share data and prolong its lifespan.

#### 2.6.1.1 [Locate the stage and direction of data sharing in the data life cycle](#)

One of the most popular data life cycle models was developed by Michener and Jones (2012) which includes 8 steps (Figure 2.3): (1) Data planning: deciding why, how, who, what, when, and where to collect data, as well as how to manage it (five “W” and two “H”); (2) Data collection; (3) Data quality assurance and control (QA/QC): approaches are adopted to ensure and control data quality; (4) Data description: the five “W” and two “H” are described clearly in metadata (Michener, 2006; Fegraus et al., 2005; Jones et al., 2001); (5) Data preservation: data is stored in repositories; (6) Data discovery: new usage and value of old data are discovered; (7) Data integration: data from disparate studies and disciplines are integrated; (8) Data analysis.

This model applies to both traditional science and citizen science: it provided the basis for Wiggins et al. (2013) to develop guidance for data management in citizen science projects. Although represented as a logical cycle, the steps in the data life cycle do not necessarily follow a fixed order and can happen in any number of different sequences depending on specific research needs (Michener & Jones, 2012; Wiggins et al., 2013). For example, most scientific projects that need to collect new data typically start from step 1 to 5 (i.e., plan, collect, assure, describe, and preserve) and can then jump to step 8 (i.e. analysis), while synthesis or meta-analysis study can start at step 6 (i.e., discover) (Michener & Jones, 2012).

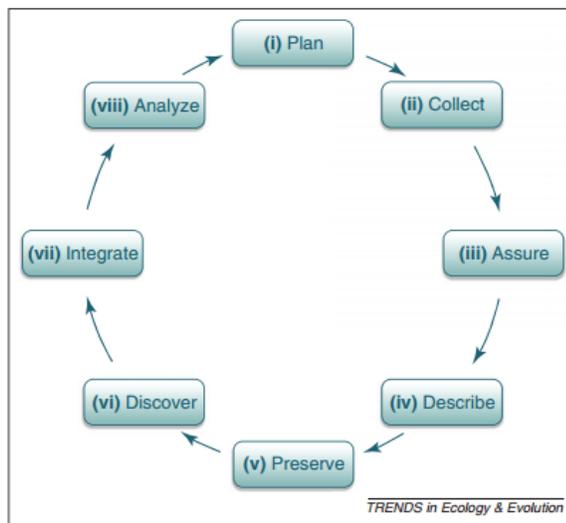


Figure 2.3 The data life cycle cited from Michener and Jones (2012).

Rüegg et al. (2014) modified the data life cycle and make it 9 steps (Figure 2.4) by adding one step of “Analyze” before “Describe”. They suggest categorizing the now 9 steps into three groups: traditional project, data re-use, and closing the data life cycle. A traditional project usually progresses through the stages of planning, data collection, QA/QC, and data analysis. Most commonly, the data is stored in a dataset on the project

researchers' computer(s). After the results of the data analysis are published, the data life cycle will likely terminate if the dataset remains private and no plan for sharing exists. The stages of preserve/publish and describe/document are necessary to prolong the life of data and encourage the data life cycle to be completed (Rüegg et al., 2014). Without appropriate digitization, documentation, and preservation, data is unlikely to be discovered and reused by others. Data discovery and integration—the following stages—are unlikely to happen.

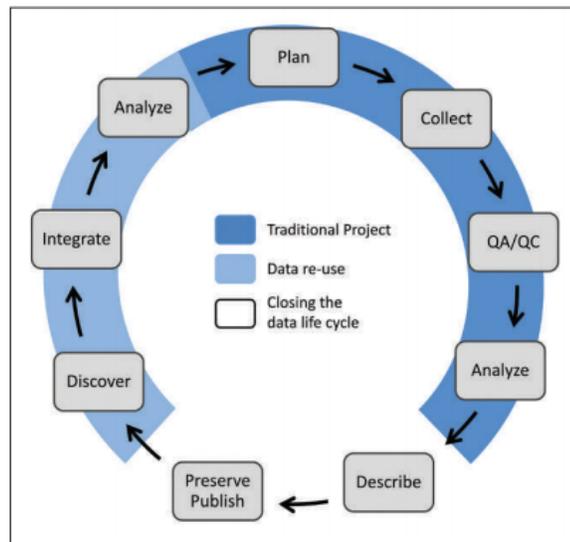


Figure 2.4 The data life cycle cited from Rüegg et al. (2014).

The data discovery stage comprises two facets: first, exploring the use of existing data in conjunction with other sources of information; and second, making a research project's data available to be discovered and accessible to others (Wiggins et al., 2013). The directions of information transformation between the two facets are opposite: while the first facet concerns active information acquisition, the second facet represents active information sharing.

These two facets can be also found in the stage of data integration. For example, at the same time that data can be aggregated and integrated for a specific research project's need, the same research project can also share its data with others within a data aggregation and integration system so other people can access it. Compared with other types of local information systems, such as databases typically only built for sharing data collected by a research team, aggregation and integration systems are more likely to be where large scale data are shared by diverse data donors. This dissertation focuses on the second facet, active information sharing, reflecting the overall theme of this research (that is, sharing data with others in an information system so that the data can be discovered and used by others).

#### 2.6.1.2 Applying the data life cycle to citizen science data

By representing its different stages of life, the data life cycle model enables us to understand the nature of data itself. However, it is important to note that while Rüegg et al. (2014) named one of the stages in the cycle “traditional project,” their use of the word “traditional” is different from its use in “traditional science” as compared with “citizen science”. Table 2.1 illustrates this difference in the context of general scientific practices versus specific scientific practice (i.e., data sharing). The fundamental difference between “traditional science” and “citizen science” is that citizen science involves non-professionals' effort in research (e.g., collecting and analyzing research data). Rüegg et al.'s (2014) “traditional” refers to the data processes that most research projects include, irrespective of whether the project is a traditional science or a citizen science project.

General scientific practices		
Specific scientific practices (i.e., Data sharing)	Traditional science	Citizen science
Traditional project	Not sharing	Not sharing
Modern project	Sharing	Sharing

Table 2.1 The differences of using the word “traditional”.

The processes falling in the other two categories—closing the data life cycle and data re-use—are still not as widely adopted as conventional life cycle processes, including in both traditional science and citizen science projects. Nevertheless, compared to traditional science projects, citizen science projects are much more likely to adopt the processes of closing the data life cycle and data re-use because of their obligation to share data (Soranno et al., 2015).

To help understand the active data sharing, the data life cycle model will be utilized in Chapter 3 to select the cases that focus on certain stages of data life cycle. However, while the data life cycle model identifies stages of data life, it is insufficient for helping understand the processes of sharing data. Therefore, a second theoretical model is needed to provide an overview of the basic processes of and influential factors on data sharing.

## 2.6.2 The framework of academic data sharing

Fecher, Friesike, and Hebing (2015) developed a cross-disciplinary framework of academic data sharing from primary researchers’ points of view (Figure 2.5). This framework provides a somewhat comprehensive overview of the fundamental

components included in data sharing processes: data donor, research organization, research community, norms, data infrastructure, and data recipients. This framework helps better map how the data sharing challenges introduced in the background literature could influence data sharing practices.

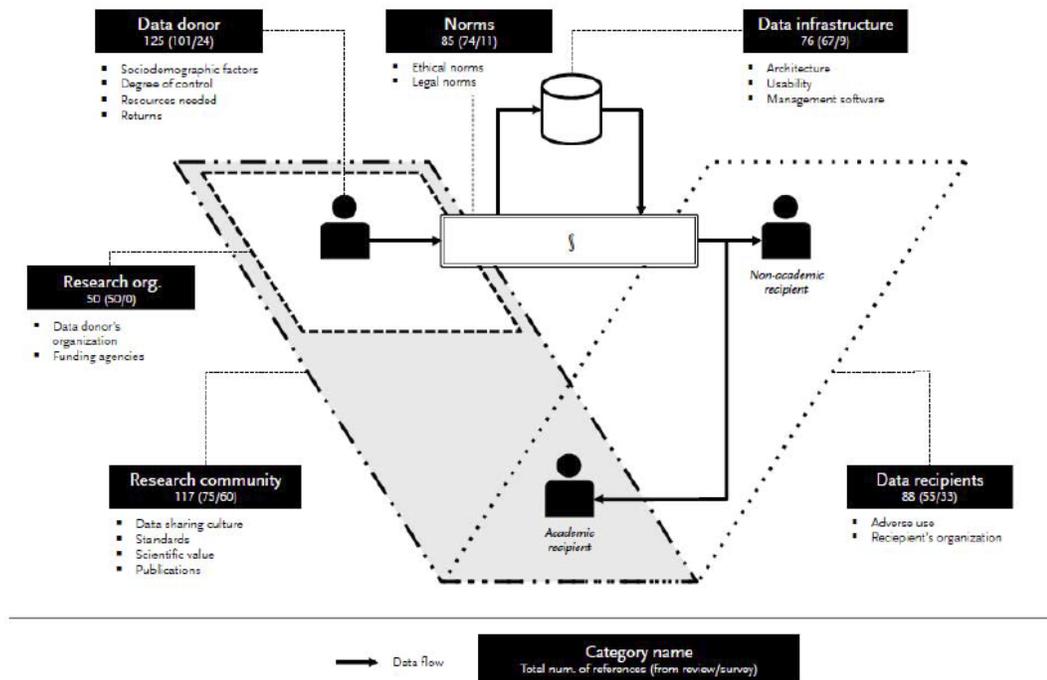


Figure 2.5 Framework for academic data sharing, cited from Fecher et al. (2015).

### 2.6.2.1 What makes data sharing happen?

#### *Data donor*

Data donor refers to the individual researchers who are responsible for collecting data (Fecher et al., 2015). Fecher et al., (2015) summarized four personal factors that could influence data sharing motivation and behavior. The first is their socio-demographic characteristics, such as nationality, age, and seniority of career. The second factor is their perception of how much they can control the usage of data after it has been shared.

Thirdly, they consider the resources they need to implement to make data sharing decisions, such as whether they have enough time, skills and knowledge, and money to finish the data sharing work. The fourth factor is the recognition and benefit they can get from sharing their data, for example whether their data shared online can be recognized on par with that published in peer-review journals.

### *Research organizations*

Research organizations, for example an affiliated organization or funding agency, are the most relevant organizational entities to the data donor (Fecher et al., 2015). These organizational entities are the chief source of external factors such as organizational data sharing policy and culture that have strong influence over shaping an individual researcher's data sharing behavior. For example, if a research institution issues a policy to require its employees (the individual researchers) to share their data in a certain repository, employees are mandated to follow this policy. In the case of funding agencies, agencies that prefer or require a detailed data management plan, including data sharing strategies, in the grant proposal will influence researchers to create such plans. These policies drive individual researchers to pay more attention to sharing data online or in an accessible repository in addition to publishing it in traditional peer review journals.

### *Research community*

Unlike an organization that usually comprises multiple structured groups characterized by different levels of power, status, and prestige (Hogg & Terry, 2000), a community in general emphasizes a much looser networking of individuals and groups who share

similar interests or goals. A strict hierarchy is not necessary for forming a community. The research community in Fecher et al.'s (2015) framework refers to a collection of individual researchers and organizations in academia who share a culture that distinguishes them from non-academic communities.

Fecher et al. (2015) summarized four high level factors conceptualized in the research community that may affect data sharing practices. The first is data sharing culture, which can be different across research disciplines. For example, compared to the natural sciences, such as biology and astronomy, sharing data in the social sciences, such as psychology and anthropology, is much less popular and seldom encouraged. Fecher et al.'s (2015) second factor is standards, such as to what degree the standards of sharing data (e.g., agreed common or united data format) are built. Varying standards could lead to confusion in preparing the data researchers want to share and choosing the data sharing tools they want to use (Linkert et al., 2010). The third factor is scientific value. Advancing science is the most well know value shared by researchers in academia. Being aware that data sharing can enhance scientific progress helps researchers to better understand the value of sharing their data and thus motivates them to share data. The final reason is publications, the primary currency in academia. Besides funding agencies, some peer review journals also have started to require authors to submit and share their datasets. This kind of journal policy is not only accepted by many researchers (Huang et al., 2013), but also provides more powerful motivation than the data sharing required by funding agencies (Enke et al., 2012).

### *Norms*

The important influence of norms on determining human behavior has been recognized for a long time (e.g., Sherif, 1936; Pratt & Rafaeli, 1997). “Group norms” refers to legitimate, socially-shared standards that affect how people perceive and interact with other people (Bettenhausen & Murnighan, 1991; Flynn & Chatman, 2002). Fecher et al. (2015) described two norms that influence data sharing behavior. First, ethical norms, such as confidentiality and the potential of information that identifies individuals make researchers hesitate to share data. Second, legal norms such as copyright mechanisms help solve issues of data ownership and use. In order to protect the ownership of data and facilitate data sharing, certain types of licenses, such as Creative Commons licenses (Creative Commons, n.d.), have been developed so that data donors retain copyright while allowing others to make legal use of their data without users needing to seek permission.

### *Data recipients*

Adverse use of the data is just one important concern researchers may have about data recipients (Fecher et al., 2015). Researchers might not want their data to be used commercially or to be used to publish a paper before they can, or they may not trust that data recipients can correctly interpret and re-use their data. In addition, some researchers might be afraid they failed to find mistakes in their data when the data recipients succeeded or that the intent of the recipient in using the data might be incongruent with their vision for its use. A further concern of sharing data may arise from the data recipients’ organizational affiliation: whether the data recipients’ lab facilities allow them

to practice good data management and maintenance, for example. (Fernandez, Patrick & Zuck, 2012).

### *Data infrastructure*

The concept of infrastructure in Fecher et al.'s (2015) study is narrower than the concept adopted in this dissertation (i.e., relational and sustainable). Instead of using the rational concept of infrastructure (Starr, 1999) and considering the human infrastructure of data infrastructure (Lee et al., 2006), infrastructure in Fecher et al. (2015)'s framework refers to all the technical infrastructure used to store and retrieve data.

Fecher et al. (2015) summarized three factors related to technical infrastructure. The first is architecture: how well the design of the infrastructure allows the data to be accessed, stored, and protected, and how data quality is assured and controlled could influence whether the researchers choose to share data. The second factor is usability. As not all researchers have a technology background, whether the data sharing infrastructure is easy to use or whether there is enough technical support could affect data sharing. The third one is the management system concerning data documentation and metadata standards (Axelsson & Schroeder, 2009; Linkert et al., 2010; Tenopir et al., 2011); there are still a great number of issues surrounding the building of clear standards (Acord & Harley, 2012).

#### 2.6.2.2 Mediators in a data sharing framework

There is little existing research that focuses on the issues of data sharing in a comprehensive manner: Fecher et al.'s (2015) team are one of the first, and among the

only, to address this problem. In their framework, the two components of norms and data infrastructure connect data donors—equivalent to “data provider” in this dissertation—and recipients—or “data users.”

However, in this dissertation norms and data infrastructure are not sufficient for illustrating what data sharing mediators do. Data sharing mediators as collective entities include people, organizations, networks, arrangements, culture, norms, and technical demands in order to develop and maintain the knowledge infrastructure to support communication, data, information, and knowledge exchange between data donors and data recipients (Lee et al., 2006).

There is limited understanding about what transpires between data providers and data users. In other words, what exactly data sharing mediators do to connect data providers with data users across research and public communities to make data sharing possible. Therefore, this dissertation considers replacing the two components of norms and data infrastructure with knowledge infrastructure. What the data sharing mediators’ data sharing practices are within the knowledge infrastructure is not clear, but will be studied in this dissertation.

### [2.6.2.3 Applying the data sharing framework in citizen science data sharing](#)

The growing popularity of citizen science reflects the increasing importance of the data collected and shared by non-professionals. It is imperative for data sharing related studies to not only consider data contributed by researchers, but also by non-

professionals, since a more comprehensive view of data sharing must include both points of view. A data sharing framework should include not only non-professional data users, but also non-professional data creators/providers so that the level of data sharing by both researchers and non-professionals is on par.

Unlike researchers, non-professional data creators/providers are not usually affiliated with a research organization. When they contribute data to research by participating in a citizen science project, they can be considered a special group of individual data providers that temporarily belong to the research community; they are citizen scientists. Others who are not citizen scientists, such as stakeholders and policymakers, still belong to the public community. The white area of the right trapezoid in the graphic representation of Fecher et al.'s (2015) framework (Figure 2.5) can be labeled as public community.

Factors that influence data sharing by non-professionals may be similar to those of the researchers, but it is more likely that they will differ. Although it is possible that non-professionals share some socio-demographic characteristics with researchers, the contexts in which non-professionals are placed are diverse and different from researchers.

With regards to technical infrastructure, non-professionals and researchers could use the same or a different data infrastructure to share data. Therefore, they might encounter similar or different technical and usability issues, depending on what social situations (i.e., with whom and for what reason) exist when sharing the data.

#### 2.6.2.4 Applying the data sharing framework to different levels of data sharing

In order to understand a more comprehensive view of data sharing, the different levels of data sharing—individual-level and collective-level—from Fecher et al.'s (2015) framework should be examined. Their framework concerns data sharing by individual-level researchers, but does not address data shared by groups of researchers or citizen scientists, a research organization, a citizen science project, or any other kinds of group. In the age of big data, many scientific disciplines have evolved to be data intensive, therefore collective-level data sharing is very common and is considered more efficient than individual-level sharing.

The difference between collective-level and individual-level data sharing can be examined from at least two aspects: the number of human actors as well as their identities. As for the number of human actors, only one individual takes the responsibility of sharing data under the condition of individual-level sharing. But under collective-level sharing, there could be more than one individual who is responsible for sharing data. Similarly, the identities of human actors in terms of individual-level sharing is simple: s/he represents him/herself as an independent researcher when making decisions about data sharing. S/he shares data on his/her own behalf and is very likely to be the creator or curator of the data. In terms of collective-level data sharing, no matter how many individual human actors participate in sharing data, the decision to and act of does not belong to any specific individual, but to a group of individuals (e.g., a research

group, an organization, a community). These individual human actors share a collective-level identity.

Therefore, a data provider can be either an individual-level or a collective-level data provider. A collective-level data provider could share a similar context with individual-level data provider if they belong to a same group. For example, the individual-level data provider is affiliated with the collective-level data provider (i.e. from researchers' perspective) or is involved in the collective-level provider (i.e. from non-professionals' perspective). Whether this context influences them in a similar or different way is unknown as individual-level and organizational-level behavior are related but different. The interrelationships between an individual-level data provider and collective-level data provider are important for understanding their contexts.

As discussed earlier, whether researchers and non-professionals share data using the same or different data infrastructures and how this would influence their data sharing, the same questions are asked for the individual-level data providers and collective-level data providers. However, there are at least two different conditions that need to be considered. The first condition is that an individual-level provider is independent from a collective-level provider. The data held by the individual-level data provider does not overlap with the data held by the collect-level provider. Under this condition, the individual-level and the collective-level data providers can be two independent providers when they share data using either the same or different data infrastructure.

The second condition is that the individual-level data provider has already shared his/her data with the collective-level data provider, but the individual-level data provider is not necessarily affiliated with the collective-level data provider. For example, when the individual-level data provider is a researcher in a human gene research lab, after s/he collects the data, s/he is required to share those data with the lab, university, or global human gene database (e.g., GenBank). Similarly, a citizen who collects butterfly observation data that includes the observation date, location, and a beautiful photograph of the butterfly who then s/he shares this data with a citizen science project by uploading them to the project database is not affiliated with collective-level provider.

In both examples, when collective-level data providers share data they held for individuals somewhere else online (i.e., a different platform), the individual-level data providers' data would be shared by the collective-level provider. Then the individual-level data providers do not need to repeat this sharing by themselves, unless they want to share their data with a different platform to the one that the collective-level data providers choose. Under this condition, the collective-level data providers become the local data infrastructure adopted by the individual-level data providers to share their data. In this scenario, the collective-level and the individual-level data providers can influence each other, and, at the same time, can be affected by other data infrastructure that they each choose to share their data with.

Understanding the interrelationships between individual-level data providers, collective-level data providers, and data infrastructure within a knowledge infrastructure is

important for understanding the contexts of data sharing. However, the current framework developed by Fecher et al. (2015) is not comprehensive enough to study both collective-level and individual-level data sharing practices as well as the complex structure of knowledge infrastructure. Within this dissertation, another theoretical framework to identify individual-level and collective-level data sharing practices and understand their interrelationships is needed.

This framework of data sharing will be returned to in the Chapter 3 in order to guide the initial analysis of data sharing processes in the selected cases and their key components.

### 2.6.2.5 The framework of data sharing processes

Based on the above discussion, a first version of this doctoral research’s theoretical framework for data sharing components and processes is created (Figure 2.6).

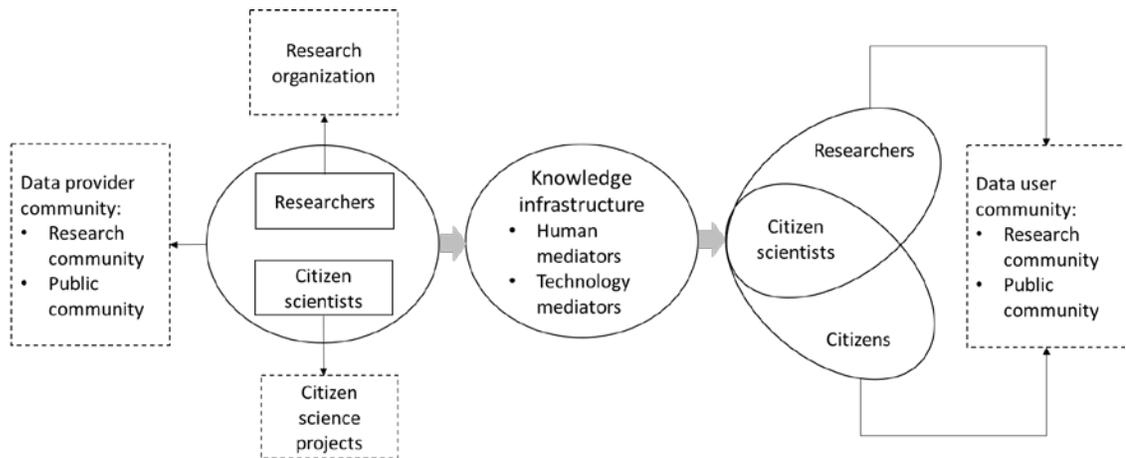


Figure 2.6 The framework of data sharing processes for this doctoral research.

### 2.6.3 The model of microfunctions of institutional logics

The discussion above has made it clear that 1) data providers can be either researchers or non-professionals; and 2) data providers can be at either an individual-level or collective-level. It is important to understand the interrelationships between individual-level and collective-level data providers that reflect the contexts of data sharing (i.e., the social situations in which they are located), because these interrelationships and contexts may have significant effects on data providers' motivation, decision, and behavior, as well as on the development and maintenance of the knowledge infrastructure by human mediators of data sharing (Bietz et al., 2010). Therefore, this dissertation needs a theoretical model to help understand these interrelationships.

The metatheoretical model of institutional logics was developed for analyzing the interrelationships between individuals, organizations, and institutions (Figure 2.7) (Thornton, Ocasio, & Lounsbury, 2012). This model explains the cross-level effects (i.e., availability and accessibility) of macro-level institutional logics on organizations (meso-level) and individuals (micro-level) (Thornton et al., 2012). An institutional logic is defined as “the socially constructed, historical patterns of cultural symbols and material practices, including assumptions, values, and beliefs, by which individuals and organizations provide meaning to their daily activity, organize time and space, and reproduce their lives and experiences” (Thornton et al., 2012, p. 2; also see Thornton & Ocasio, 2008).

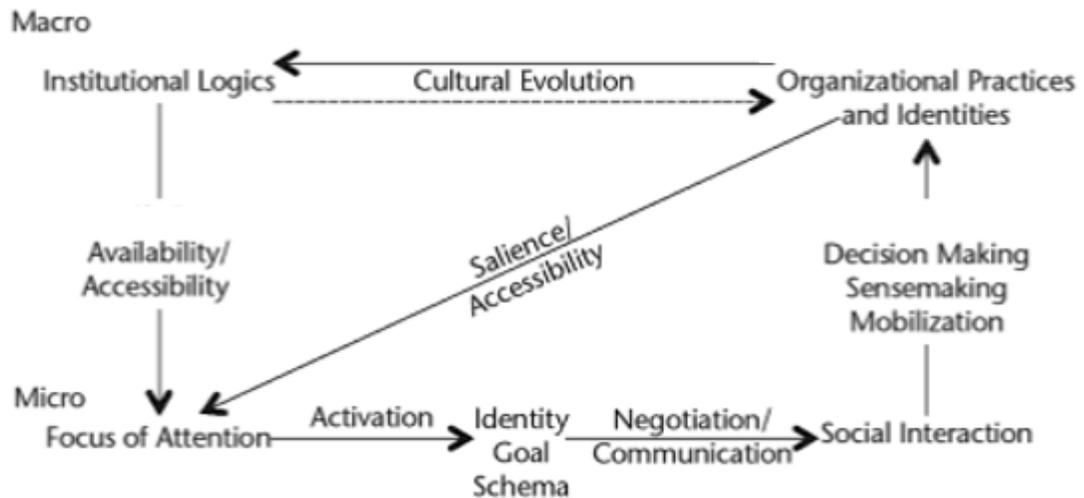


Figure 2.7 A cross-level model of institutional logics combining macro-micro and micro-macro, cited from Thornton et al. (2012).

This dissertation adopts it to understand the influence of the interrelationships between individuals, organizations, and institutions among different data providers and mediators of data sharing. This approach aids in the investigation of data sharing contexts and the complex environments of knowledge infrastructure.

The interrelationships between individuals, organizations, and institutions are reflected by how individual and organizational actors are influenced by their social situations in an interinstitutional system (Thornton et al., 2012). In this dissertation, the individual-level and collective-level data providers and the human mediators of data sharing correspond to individual and organizational social actors. Collaborative research data sharing at both individual and collective levels across research and public communities happens in an interinstitutional system.

The model of microfunctions of institutional logics focuses on three characteristics of social actors, being social identities, goals, and schemas (Thornton et al., 2012). A social actor can have multiple social identities and goals that act as motivators for the actor and are embedded within alternative institutional logics (Thornton et al., 2012). Social identities help actors recognize the roles that they play and the work do in different situations, while goals guide social actors' cognition and current actions, as well as the plan for and expectations of the future (Thornton et al., 2012). Schemas are "learned, organized cognitive structures that shape attention, construal, inference, and problem solving" (Thornton et al., 2012, p. 88; see also Nisbett & Ross, 1980). Institutional logics are one important source that help social actors develop top-down cognitive and knowledge structure (Thornton et al., 2012). For different actions and goals, or for different logics, there could be different schemas (Cheng & Holyoak, 1985; Thornton et al., 2012).

Multiple social identities, goals, and schemas are not equally available and accessible to actors in different social situations: some of them are more accessible and more likely to be activated (Thornton et al., 2012). Institutional logics, together with social structures and practices, influence social actors' focus of attention that make their identities, goals, and schemas available, accessible, and activated in diverse social situations (Thornton et al., 2012). Social structure and focus of attention are described briefly in the following.

Social structure emphasizes the importance of relationships between different social actors. It can be abstracted from the concrete population and its behavior in society, and

refers to the pattern, network, or “system” of relationships between social actors in their capacity of playing roles (i.e., social identities) relative to one another (Nadel, 2013). These relationships are social or institutionalized, indicating that social actors are influenced by one another, and have some consist and constant attributes that differentiate social actors’ acts from single or disjointed acts (Nadel, 2013).

Focus of attention can be shaped by both top-down (i.e., goal or schema driven) and bottom-up (i.e., stimulus driven) processes (Ocasio, 2011). Intentional and sustained allocation of cognitive resources—attentional engagement—is necessary to guide problem solving, planning, sensemaking, and decision making (Ocasio, 2011).

Institutional logics help determine the focus of attention, allocating how many cognitive resources from whom (i.e., social identities) to focus on what problems and solutions in what ways (i.e., goals and schemas) (Thornton et al., 2012; see also Ocasio, 1997; Thornton and Ocasio 1999; Thornton, 2004). Although the focus of attention describes individual cognitive processes, the concept can be applied to organizations and institutions (Thornton et al., 2012).

The available, accessible, and activated identities, goals, and schemas then shape social interaction that can be both material and symbolic and within which negotiation, exchanges, and communication among different social actors is central (Thornton et al., 2012). Social actors who encounter each other and participate in a social interaction (e.g., cooperation and collaboration) must, to some degree, share in the focus of attention on the contents of interaction.

However, the actors do not have to be embedded in the same culture or institutional logics. The theory of dynamic constructivism can be adopted to explain how multiple institutional logics are available, accessible, and activated (Thornton et al., 2012; see also in Brett, 2010). Based on this theory, different identities, goals, and schemas can be activated in different cultures or institutional logics when social actors encounter the same situation (Thornton et al., 2012). Organizational practices and identities are then formed by these social actors under three mechanisms: decision making, sensemaking, and mobilization (Thornton et al., 2012). Decision making is the core of understanding organizational processes (e.g., decision rules, performance programs, and routines) and outcomes (e.g., structure and design) (Thornton et al., 2012; see also Barnard & Simon, 1947; March & Simon 1958; Cyert & March, 1963). The process of sensemaking claims that social actors “turn circumstances into situations that are comprehended explicitly in words and serve as springboards for action” (Thornton et al., 2012., p. 96; see also Weick, Sutcliffe, & Obstfeld 2005). Finally, the process of mobilization claims that social actors gain symbolic and material resources and motivate human actors to accomplish the collective-level goals for the groups (Thornton et al., 2012).

#### 2.6.3.1 Applying the framework of microfunctions of institutional logics in scientific data sharing

Given the discussion of “traditional project,” “traditional science,” and “citizen science” above, there seems to be at least two general types of historical patterns of data management and scientific practices. The pattern of the “traditional project” in which the stages of closing the data life cycle and data re-use are not included (Rüegg et al., 2014)

is different from that of the “modern project” in which these stages are included.

Similarly, the pattern of “traditional science” in which the data creators are scientists and experts is different from that of “citizen science” in which the data is collected and/or analyzed by non-professionals.

These different patterns could signal distinctive institutional logics that might influence the focus of attention of the data providers and human mediators of data sharing. Their identities, goals, and schemas become accessible, available, and activated in different instances of data sharing and then go on to shape their social interaction regarding data sharing. Following this, their organizational-level data practices and identities are formed by data providers and human infrastructure via decision making, sensemaking, and mobilization.

The institutional logics that underlie data sharing practices are mostly comprised of collaborative efforts made by data providers and data sharing mediators. Although the focus of this dissertation is not to identify specific institutional logics in scientific practices and data sharing, having the guidance of the microfunctions of institutional logics model is helpful for understanding the interrelationships between individuals, organizations, and institutions involved in data sharing, as well as understanding the collaboration efforts between data providers and mediators.

There is no uniform procedure to identify institutional logics. Identifying institutional logics needs more information than fuzzy historical patterns of data management and

scientific practices. This dissertation focuses on the concept of identity to gain an initial understanding of the underlying institutional logics of data sharing. The concept of identity has been used in previous research on institutional logics to explain the conditions of organizations and institutions, and the identity of social actors is able to embody the institutional logics (Thornton et al., 2012; see also Thornton & Ocasio, 1999; Rao, Monin, & Durand, 2003).

Compared with the other two characteristics of the social actors (i.e., goals and schemas), social identities are relatively more visible and easier to obtain. Although a group (e.g., organization, community) can experience changes in identity at different stages of its development, ascertaining a clean-cut collective identity is still the most important prerequisite for establish a group as a collective-level social actor. For example, an organization's collective-level identity indicates how it understands itself with regards to who it is and how it is uniquely unlike other organizations (Tyworth, 2014). This identity is usually described clearly and openly offline and/or online. The importance of locating the collective-level identity is also the same for building a community or other types of group.

#### 2.6.3.2 The framework of data sharing contexts

Based on the discussion applying Thornton et al.'s (2012) model of microfunctions of institutional logics to scientific practices and data sharing, the first version of the analytical framework of data sharing contexts for this dissertation is created (Figure 2.8).

Social interaction between data donors and human mediators of data sharing reflect data sharing processes.

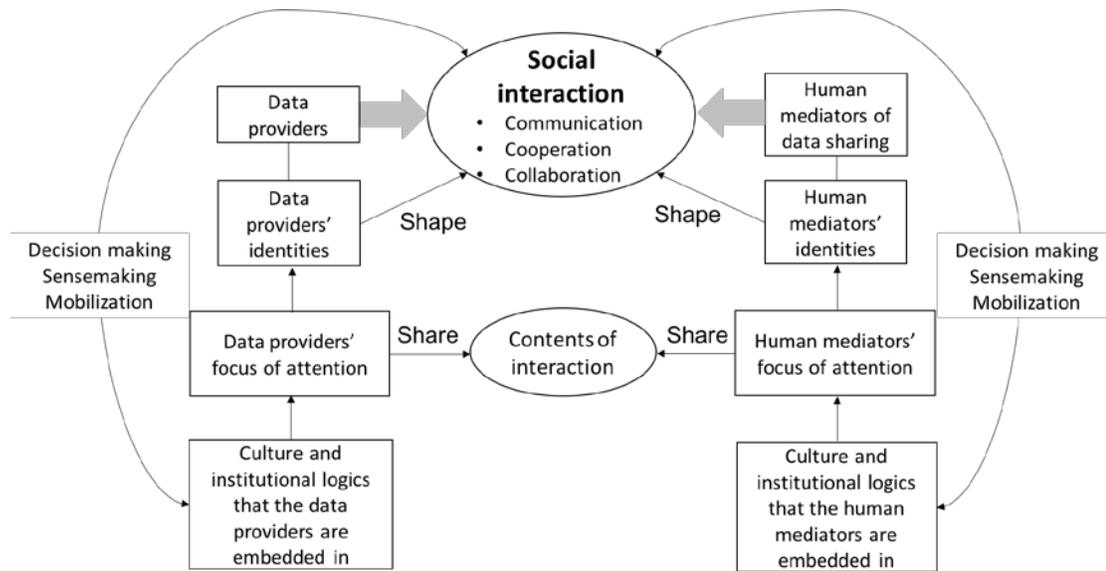


Figure 2.8 The framework of data sharing contexts for this dissertation.

#### 2.6.4 The model of identification

This dissertation research takes advantage of organizational studies by adapting an organizational identity model to fit the analysis of the collective-level and individual-level identities in data sharing (Figure 2.9). In organizational studies, identity and identification are “root constructs,” that is, each entity should have a sense of who or what it is, who or what other entities are, and how the entities are associated (Albert et al., 2003; Ashforth et al., 2008). Identities help individual human actors gain a sense of the social landscape by situating different entities, and identification embeds individual human actors within the relevant identities (Ashforth et al., 2008). Although identity can be understood as personal identity referring to the unique sense of self (Postmes & Jetten, 2006), this dissertation focuses on the social identities that are “rational and comparative”

(Tajfel & Turner, 1986), such as “who are we?” or “who am I related to the group I belong to?” (Ashforth et al., 2008).

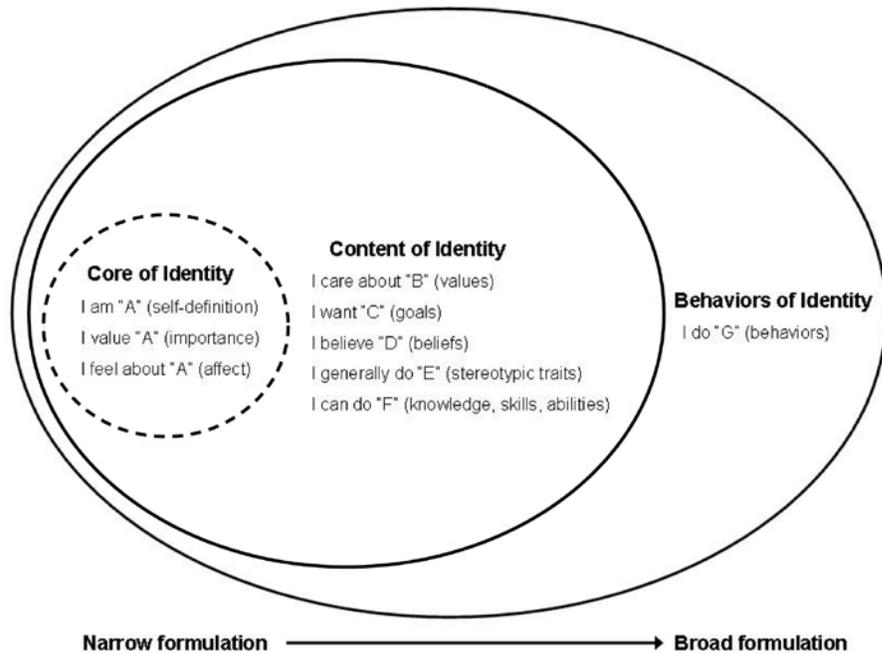


Figure 2.9 A fuzzy model of identification cited from Ashforth et al., (2008).

Identification can be considered as the process of self-defining and identity formulations. It is also “the perception of oneness or belongingness to some human aggregate” (Ashforth & Meal, 1989, p. 21). Ashforth et al., (2008) depicts a fuzzy model of identification, ranging from narrow to broad identify formulations (Figure 2.9). At its narrow end, the core attributes of identification include “I am A, I value A, and I feel about A.” Between the narrow and broad ends, the central, distinctive, and more or less enduring, attributes of identification in organizational contexts include values, goals, stereotypic traits, and knowledge, skills, and abilities. They comprise the content of identity. Identities can, but do not necessarily include all the content attributes. At the broadest end, the attribute of identification is about behavior.

Although it is also not a necessary element of identity, it can be considered as a probabilistic outcome of identification, which can be important for individuals' self- and social-construction of identification; in other words, not only thinking and feeling their ways into identification, but also acting their ways into it (Ashforth et al., 2008; Ashforth, 2001). In this dissertation, the behaviors could be sharing data and any social interaction related to making data sharing possible. This dissertation uses this model of identification to understand human actors' identities and the social landscape and structures in which they are situated while data sharing.

#### 2.6.5 Integrate theoretical frameworks

In this dissertation, the four models/frameworks are chosen for selecting the study cases and analyzing the data sharing practices. Based on Figure 2.2, a clearer two-way relationship is presented in Figure 2.10. From left to right, data sharing processes reflect data sharing contexts. The model of data life cycle helps to understand the origin of data and identify at which stage of data life the data are shared. The academic data sharing framework shows the big picture of the academic data sharing phenomenon, indicating what components should be expected to appear in the processes of data sharing and the influential factors for each component. The components are at different levels (i.e. individual level and collective level) and they are not isolated; instead, they have close relationships with each other. The model of identification helps to identify the identities of the components in the processes of data sharing. Finally, the model of microfunctions

of institutional logics helps to explain the complex interrelationships among those components and the social and technical structures in which they are embedded.

From right to left, the data sharing contexts shape the data sharing processes. The culture and institutional logics of data sharing decide the identities of social actors who are responsible for sharing data. These social actors themselves are the human components included in data sharing, and they create other symbolic and material components that are necessary for data sharing. Because the social actors take actions to share data, data life is prolonged, and the data life cycle becomes possible.

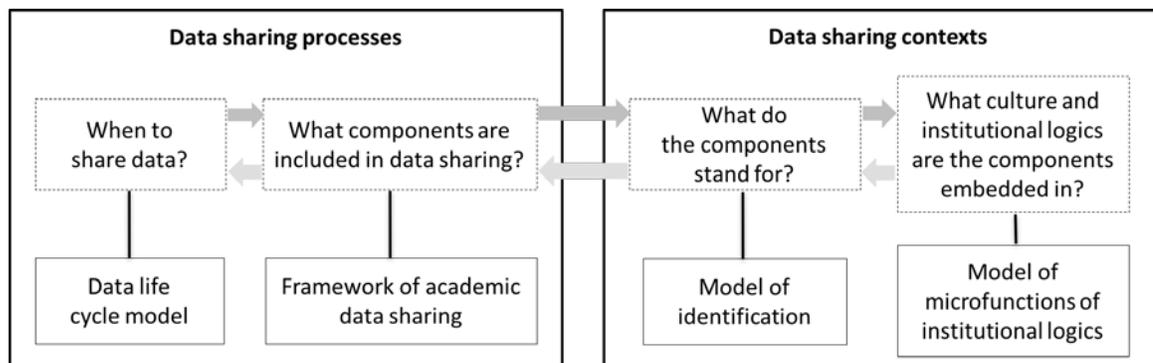


Figure 2.10 The two-way relationship among the models/frameworks.

## 2.7 Conclusion

This chapter had two major goals. The first was to introduce background literature on the definitions of data, data sharing mediators, data sharing infrastructure, and data sharing challenges. The second goal was to introduce four existing theoretical frameworks/models (i.e., the model of data life cycle, the framework of academic data sharing, the model of microfunctions of institutional logics, and the model of

identification) that could be adapted in this research for understanding data sharing contexts and processes.

The literature on the definition of data and the data life cycle model explained the nature and statuses of the key study object (i.e., data) of this dissertation. The literature on data sharing mediators and data sharing infrastructures showed the essential supporting conditions and environments provided by human and technology actors to enable data sharing. The literature on framework of academic data sharing presented different components that are involved in data sharing phenomena. Although these existing literature have made an impressive effort to understand data sharing practices and develop data sharing frameworks, they are not sufficient to answer the overarching research question (i.e., how data are shared effectively across research and public communities) for three major reasons.

First, the literature did not provide a comprehensive and in-depth understanding of the two key type of performers, being data providers and data sharing mediators, are involved in sharing data across research and public communities. They lack a systematic and comprehensive understanding of the different data providers. They demonstrated an awareness of different data from different contexts and that where the data came from could differently influence the human actors' work on developing and maintaining knowledge infrastructure. However, they did not provide a systematic understanding of the differences among data sources or on the exact influences of specific aspects of data sources on data sharing practices. These literatures predominantly focused on data

sources created by researchers and sharing these data with other researchers in the research community. However, sharing across research and public communities means sharing data sources created by researchers and non-professionals (i.e., data providers) in different contexts with both other researchers in the research community and non-professionals in the public community (i.e., data users). Therefore, these literatures could not answer the research question effectively without a more comprehensive and in-depth understanding of the differences between the data providers and the influences rooted in their differences in data sharing.

For data sharing mediators, especially for human mediators, a more in-depth understanding is needed to answer the research question. The group of human mediators is the essential part of human infrastructure that is specifically linked to the facilitation of data sharing among all functions of a knowledge infrastructure. The literature focused on understanding the human infrastructure from a holistic point of view. Combined with human infrastructure itself being less visible than other parts of a knowledge infrastructure, there is limited understanding of human mediators and their work. Therefore, the research question can only be answered if a more in-depth understanding about human mediators and the exact activities they perform to enable data sharing across research and public communities is gained.

Second, the literature provided the explanations of the different components involved in data sharing. However, it did not sufficiently acknowledge the interrelationships between these different components when considering collaboratively sharing data across research

and public communities. One important attributes of the components involved in data sharing is that they are cross-level entities (e.g., individuals, communities, organizations, institutions, etc.). Accordingly, data sharing is at different levels as well. This kind of large-scale data sharing must involve different levels of sharing, especially in terms of data sharing across research and public communities. Therefore, without understanding the interrelationships between the different components at different levels that are involved in data sharing, it is hard to understand how they collaboratively make data sharing occur across research and public communities. In addition, this is also why it is necessary for this dissertation to adopt the model of microfunctions of institutional logics and the model of identification to help understand the interrelationships between these different components at different levels.

Third, given that data sharing across research and public communities should include different levels of sharing, answering the research question requires evidence to show that data actually go through the synergetic and comprehensive processes at different levels and are shared across research and public communities in real-world cases.

However, the literature does not include an investigation to reveal these kind of data sharing processes. Therefore, this dissertation will choose and examine real-world cases to provide evidence to illustrate how data are shared effectively across research and public communities.

## 3 Methods

### 3.1 Research design

#### 3.1.1 Making sense of the research design

The overarching question of this dissertation is: *how data are shared effectively across research and public communities?* The research conducted for this dissertation investigates data sharing practices for sharing data across research and public communities. The key players are human mediators who connect data providers and data users (Wiederhold, 1992; Borgman, 2015). These human mediators constitute a human infrastructure within a knowledge infrastructure, their contributions to which build and maintain it. This dissertation will investigate how human mediators accomplish data sharing work by answering the three sub research questions:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

Answering the first question is the first step of this research. The social and technical structures of data providers have important effects on cyberinfrastructure development processes (Bietz et al., 2010). The data sharing processes studied in this dissertation can be considered the knowledge infrastructure development processes. Therefore, before examining data sharing processes, it is important to understand who the data providers are and uncover their identities. Simply put, these identities are activated and made

available by the culture and institutional logics of the organizations, institutions, or communities the data providers belong to (Thornton, Ocasio, & Lounsbury, 2012). By understanding data providers' identities, light is shed on the culture and institutional logics of their organizations, institutions, or communities, which in turn reflects the contexts of data sharing from data providers.

The second step is to answer the question: who are the data mediators? Data sharing mediators within a knowledge infrastructure include human mediators and technology mediators. Data providers make their data available to users through data mediators. Knowledge infrastructures (e.g., cyberinfrastructure) includes human infrastructure and technology infrastructure. The human infrastructure is mostly invisible or, at least, much less visible than the technology infrastructure. However, it does not mean that human infrastructure is less important than technology infrastructure (i.e., cyberinfrastructure). Without human infrastructure, cyberinfrastructure could not emerge and function well (Lee, Dourish, & Mark, 2006; Bietz, Baumer, & Lee, 2010). Likewise, this dissertation argues that without the human mediators of data sharing or if data sharing related functions of any knowledge infrastructures could not emerge and function well, data sharing processes could not even get begin. Therefore, this dissertation prioritizes revealing the critical but invisible part of data sharing mediators: human mediators.

The answers to the first and second questions indicate who (data providers, human mediators) would carry out data sharing practices (sharing data across research and public communities) and in what ways (e.g., supported by what kind of technology mediators).

Then the third step is to answer the final question by identifying what exactly are the data sharing processes. The process here refers to a series of interrelated events that are carried out over time for achieving an organizational outcome of interest (Boudreau & Robey, 1999; see in Crowston, 2000). These interrelated events include the sequential actions that are taken by the key actors (i.e., data providers and data sharing mediators) (Crowston, 2000). The process is a way to achieve this outcome (Crowston, 2000). The major organizational outcome of interest is sharing data across research and public communities.

The process itself can be considered a theory, even though the process could be very specific: the process as a theory could be revealed by describing a single performance in a specific organization (Crowston, 2000). However, “more desirably, the theory might describe a general class of performances or even performances in multiple organizations” (Crowston, 2000, p. 151). Investigating the third question in this dissertation requires analysis of data sharing practices in multiple organizations and communities.

Furthermore, Crowston (2000) also argued that “process theories provide a link between individual and organizational phenomena and a milieu for interplay between research paradigms” (p. 9). Since dissertation is oriented toward the human mediators rather than the technology mediators of data sharing, the actions in the data sharing processes will be mainly social (inter)action (i.e., communication, cooperation, collaboration) that are initiated and taken by human actors with the support of the technology mediators.

Technology mediators operate the information system (i.e., technology infrastructure)

that is set up beforehand; system developers and designers (i.e., human actors) put the mediators into the control program of the information system. In other words, these processes are initiated by and under the control of human actors. Therefore, this dissertation focuses on the social interactions among the human mediators of data sharing in order to successfully align and leverage collaborative data sharing relationships with each other so as data sharing across research and public communities takes place.

### 3.1.2 Research methods

To answer these three questions, two methods—infrastructural inversion (Bowker, 1994) and case study (Yin, 2013)—are adopted as a conceptual method and a research method respectively.

Infrastructural inversion (Bowker, 1994; Bowker et al., 2010) is a conceptual method commonly used in studying social and technical infrastructures. When the infrastructure is considered a relational and ecological concept, most parts of the infrastructure could be invisible (Star & Ruhleder, 1994; Star, 1999). For example, the development and maintenance of infrastructure are cared for by invisible human workers who are the essential parts of the infrastructure (Star, 1991). Infrastructural inversion is a methodological device used by researchers study infrastructure itself to illuminate a clear direction: inner workings of the infrastructure are brought to the foreground and its relational nature revealed in that it “emerges for people in practice, connected to activities and structures” (Bowker et al., 2010, p. 99; Lee, et al., 2006). This dissertation takes advantage of this methodological device to expose and investigate the work that has been

done by human mediators within knowledge infrastructures. These knowledge infrastructures are built for, but are not limited to, sharing data across research and public communities.

This dissertation selects two real-world cases of knowledge infrastructure built for sharing data within both the research and public communities. The case study method has a distinct advantage when researchers ask “how” or “why” questions about “a contemporary set of events ... over which a researcher has little or no control” (Yin, 2013, p. 14), a state reflected in the research in this dissertation. The overarching research question of this dissertation is a “how” question: *How data are shared effectively across research and public communities?*

Yin (2013) suggests a twofold definition of case study: (1) from the perspective of scope, a case study is an empirical inquiry that investigates a contemporary phenomenon in depth in its real-world context, especially when the boundaries between the phenomenon and context are blurred; (2) from the perspective of features, there will be many more variables of interest in a case study than merely the number of data points, with results relying on data collection and analysis from multiple sources guided by prior development of theoretical propositions. As an infrastructure is usually largely invisible (Star & Ruhleder, 1994), the visible product of knowledge infrastructure development (e.g., data repositories) (Bietz et al., 2010) is used as the entry point for selecting cases.

To add confidence to the findings of this dissertation, more than one case should be chosen by following a replication strategy (Miles, 2014; Yin, 2013). Therefore, a comparative case study design is adopted in this dissertation. The replication logic is different from that of a statistical “sampling” design; instead it is a theoretical replication that reflects the theoretical interests of researchers (Yin, 2013). The theoretical interests of this dissertation have been introduced in Chapter 2 and is reflected by two frameworks: data sharing contexts and data sharing processes.

Two real-world cases were chosen in this dissertation: (1) a large scale knowledge infrastructure named Encyclopedia of Life with its product, an aggregator repository, also named Encyclopedia of Life; and (2) a relatively small scale knowledge infrastructure, Cyber-Innovation for Sustainability Science and Engineering (CyberSEES), with its product, a citizen science project data repository, Biocubes. Both cases are located in the domain of biodiversity but represent different data sharing contexts and processes. Multiple sources of data are collected from these two cases through participant observation, interviews, documentation, and artifacts. Data collection and analysis proceeded simultaneously. For each case, all three sub questions are answered for within-case analysis. Answers are then compared for cross-case analysis.

### 3.2 Domain selection

The domain of biodiversity was selected before selecting specific cases. The word biodiversity can conjure up a boundless image of countless creatures. This domain was chosen not only because its data has become “big” and thus has more potential than ever

to bolster the advance of science (Hampton et al., 2013; Howe et al., 2008), but also because of my personal interest in and concern for this domain, led by the sixth extinction wave.

When more than 75% of species die out in a geologically short interval, paleontologists characterize this as a time of mass extinction (Barnosky, 2011). This planet is experiencing its sixth mass extinction, triggered by human actions in the past 500 years (Barnosky, 2011; Dirzo et al., 2014). The fifth one was marked by the extinction of dinosaurs, thought to have been caused by a natural disaster. The culprit of the sixth mass extinction is none other than humans. For thousands of years, humans have occupied wildlife habitats and polluted the air, water, and soil. Consequently, many species have and continue to vanish from the world we persistently dominate the world at the expense of other organisms.

Biodiversity is an essential part of our natural ecosystem, providing us with food, medicines, and industrial products, without which humans can barely sustain life (Ehrlich & Wilson, 1991; Novacek, 2008). Unfortunately, the urgency of losing biodiversity has not yet attracted enough public attention. Curry et al. (2007) conducted a survey to investigate public attitudes toward environmental issues, and the results show that among the 18 most important social problems facing the US, environmental issues are only ranked 13th, below terrorism, the Iraq war, health care, the economy, education, the quality of government leaders, social security, illegal immigrants, and family values. Among the 10 most important environmental problems facing the US, destruction of

ecosystems and endangered species ranked 2nd and 9th respectively; other problems ranked from top to bottom are global warming, water pollution, overpopulation, toxic waste, ozone depletion, urban sprawl, smog, and acid rain (Curry et al., 2007). There is still ample opportunity to increase public awareness and engagement in biodiversity issues.

One possible reason for this current situation is that people do not think protecting biodiversity is an important enough issue and that the negative effects of losing biodiversity is not as direct and immediate as that of terrorism, war, economy, and education. People are much better at focusing on short term gain than long term benefit. Furthermore, understanding biodiversity issues requires more scientific knowledge than other social problems and it is scientists, not the public or politicians, that are playing a leading role of protecting biodiversity and increasing public awareness of this important issue. Given the current seriousness of species extinction (Jenkins, 2003), an essential endeavor is curating biodiversity data in as comprehensive form as possible before it lags even further behind the speed of extinction and the development of biology. Following data curation, the next important step is to make biodiversity data easily accessible and understandable by anyone through taking advantage of the Internet and information and communication technologies. The development of knowledge infrastructures for sharing biodiversity data with the public has never been more important.

### 3.3 Case selection

The two cases were not selected at the same time. Following Wiggins's (2012) dissertation work, one case was selected first. The early stages of data collection and analysis and work-in-progress results gained shaped the theoretical replication case selection criteria for the next case.

#### 3.3.1 Theoretical sampling

The theoretical sampling criteria is developed based on the frameworks of data sharing contexts and data sharing processes (see Chapter 2). The framework of data sharing contexts was developed based on the model of data life cycle (Rüegg et al., 2014) and the framework of academic data sharing (Fecher et al., 2015). The framework of data sharing processes was developed based on the model of microfunctions of institutional logics (Thornton et al., 2012) and the model of organizational identities (Ashforth et al., 2008). Variations between data sharing contexts and processes are considered for sampling in this dissertation. Data sharing contexts include both offline and online contexts (i.e., online environments), but because offline sharing contexts are much less visible when selecting cases than the online, online sharing contexts are used to select the cases. Online sharing contexts refer to the online environments where the data providers share their data. The two cases selected in this work should represent different variations on the theoretical sampling criteria.

Case selection started with taking a knowledge infrastructure built for sharing research-level data across research and public communities and identifying a product of it, that

being a large-scale biodiversity aggregator repository. Based on the model of the data life cycle (Rüegg et al., 2014), data aggregation and integration systems are most likely to be the place (i.e., online environment) where large-scale data from diverse data providers are shared. These kind of systems in a specific knowledge domain could provide the gateway for discovering what kinds of data providers are in this domain. Therefore, this dissertation chose its first case: a knowledge infrastructure which has produced a successful large-scale aggregator repository in the domain of biodiversity. This aggregator repository is called Encyclopedia of Life (EOL) (eol.org). Its targeted data users include any members from both research and public communities.

The first round data collection and analysis of this case indicated that the online environment (i.e., data sharing context) provided by EOL is mainly for diverse collective-level data providers to share data, rather than individual-level data providers (i.e., individual data creators). This preliminary result is critical for understanding the different types of existing data sources in biodiversity domain, which will be reported in the findings section in Chapter 4. The collective-level data providers share data with EOL by building the formal data sharing partnerships with EOL. Building these partnerships allows EOL to aggregate this data and present them on the EOL platform.

However, individual-level data providers cannot share data directly on EOL platform: they have to share data with one of EOL's collective-level data providers first. Their data might then be shared on EOL via the partnerships. Therefore, the processes of sharing data by individual-level providers could not be revealed through the first case and,

therefore, understanding of data sharing practices is not comprehensive. Thus, the second case should be able to provide an example of an online environment (i.e., data sharing context) for supporting data sharing by individual-level data providers, so that the processes of data sharing by these providers can be revealed and analyzed.

Besides representing a variation from the frameworks of data sharing contexts and data sharing processes from the first case, another important sampling criterion for the second case is that its product, the data repository, should be one of the collective-level providers on EOL. This is so that the data sharing processes from the individual-level providers to a collective-level provider, and the data sharing processes from the collective-level provider to EOL, can be linked and the comprehensive data sharing processes can be studied.

Prior to selecting the second case for this dissertation, EOL had more than 270 collective-level data providers. Among these collective-level data providers, this dissertation decided the second case's data product should belong to one type of collective-level data provider: citizen science initiatives. The citizen science data providers on EOL exist in the form of biodiversity data repositories with the functions of an online community and social network site.

There are two major reasons for choosing this type of collective-level data provider. First, as mentioned in Chapter 1, the current mainstream culture, behaviors, and practices of researchers lack a strong ethical impetus for sharing data (Soranno et al., 2015); but

citizen science, as a relatively new scientific practice, has a strong ethical impetus for sharing data openly with both researchers and non-professionals. Understanding the data sharing processes from a citizen science project repository to an authoritative biodiversity data aggregator repository can provide a valuable example of data sharing across research and public communities. These examples can be used as references for developing both mainstream research and citizen science data sharing practices in the future.

Second, data sharing practices by researchers are well understood (e.g., Kowalczyk & Shankar, 2011; Fecher et al., 2015; Soranno et al., 2015). However, there is relatively limited understanding about the sharing practices of research data collected by non-professionals. Choosing citizen science collective-level data providers provides a research opportunity to enhance the current understanding of these sharing practices.

Based on the theoretical sampling criteria described in previous paragraphs, this dissertation chose its second case: the CyberSEES project, a cyberinfrastructure project in which a citizen science project named Biocubes was developed as the vehicle for creating citizen science data that is be ready to share between research and public communities. CyberSEES recruits non-professionals to collect biodiversity observation data. Given its limited resources, CyberSEES adopted an existing technology platform called iNaturalist ([inaturalist.org](http://inaturalist.org)) to manage data, rather than developing a brand new data repository. CyberSEES created a page, called Biocubes, on iNaturalist where project data is stored. Biocubes participants share their data in the iNaturalist platform and link their data to Biocubes.

iNaturalist is a citizen science data provider on the EOL platform and allows citizen science project participants to directly upload data to their platform. Citizen science projects can use iNaturalist to manage this data. Similar to the first case, the Biocubes project page (i.e., the Biocubes data repository) on iNaturalist can be considered the visible infrastructure product of the knowledge infrastructure. Targeted data users of iNaturalist include members from both the research and the public community (Loarie, 2016b), which is consistent with the data users targeted by CyberSEES.

The two cases, EOL and CyberSEES, were selected based on what they represent in different dimensions of the theoretical sampling criteria. In addition, the two cases are not independent from one other: their online environments are linked by the established formal data sharing partnership between EOL and iNaturalist, which allows the data shared in the CyberSEES online environment (i.e., the iNaturalist and the Biocubes project page) to be transferred and shared in the online environment provided by EOL (i.e., the EOL data aggregator repository). These two online environments support different levels of data providers (i.e., support data sharing by individual level data providers vs. support data sharing by collective data providers) and scales of the data (i.e., collecting biodiversity observation/occurrence data vs. aggregating knowledge about life on Earth). Table 3.1 summarized the two cases selected based on theoretical sampling criteria.

<b>Case selection criteria</b>		<b>Cases</b>	<b>Encyclopedia of Life (EOL)</b>	<b>The CyberSEES project (CyberSEES)</b>
<b>(Online) data sharing contexts</b>	<b>Different levels of data providers</b>		For collective-level data providers	For individual-level data providers
	<b>Different scales of online environments</b>		Aggregating and presenting different types of biodiversity data that constitute the knowledge about life on Earth	Collecting and presenting biodiversity observation/occurrence data
	<b>Sharing platforms</b>		EOL	iNaturalist/Biocubes project page
<b>Data sharing processes</b>			Collective-level data providers build data sharing partnerships with EOL	Individual-level data providers participate in the citizen science project

Table 3.1 Two cases selected based on theoretical sampling criteria.

### 3.4 Selected cases

#### 3.4.1 Encyclopedia of Life

EOL knowledge infrastructure's product is a large scale aggregator repository in the domain of biodiversity. Previous research also describes it as an open access online database (Parr et al., 2014) and a content curation community (Rotman et al., 2012). It brings together comprehensive information about all named life on Earth (e.g., 3.5 million distinct pages for taxa with more than 1.3 million of these having detailed content) into the same online place and is freely accessible to anyone with an Internet connection (Parr et al., 2014). Information includes species names, geographical distribution, maps, images, habitat descriptions, their importance for humanity, and other descriptive data presented by various types of media, such as text, images, video, sounds, maps, classifications, and more (Wilson, 2003). The wide variety of data are aggregated

from a broad array of disparate data collections to create a high quality, comprehensive resource that supports research, education, and public awareness of critical issues in biodiversity (Parr et al., 2014). These data collections are shared by diverse data providers worldwide, called EOL Content Partners.

The EOL repository has experienced a two-phase development. The first phase built infrastructure for aggregating and curating basic content with the original website launching in 2008 (Schopf et al., 2008). The second phase focused on making improvements to create a more engaging, personal, accessible, and internationalizable repository (Parr et al., 2014). The latest version of EOL (i.e., Version 2) was released in September 2011 (Parr et al., 2014).

The extensive range of targeted data providers and users on EOL ensures that they are located in various social situations. EOL provides a large-scale online environment in which there might be more than one biodiversity data sharing institutional logic. Within the data life cycle, the data shared on EOL are at the stage of being aggregated and integrated so that data users can find and use them.

### 3.4.2 The CyberSEES project

Starting in January 2015, Infrastructure and Technology Supporting Citizen Science Data Usage and Distribution for Education and Sustainability is an NSF funded cyberinfrastructure project that has adopted the name of its higher level program, CyberSEES, as the project nickname. Currently citizen science data sharing is poorly

supported, limiting the achievements of not only public engagement in science, but also scientific studies that could benefit from access to these data (CyberSEES project description, 2015). CyberSEES project members developed the Biocubes project as the vehicle for creating fresh citizen science data and studying infrastructure development and design for sharing this data (CyberSEES project description, 2015).

The concept of a biocube comes from esteemed nature photographer David Liittschwager (Liittschwager, 2012). Liittschwager built a 12-inch green metal frame, and took it into nature where he placed it in different environments from deep in a forest to shallow in a sea, from Costa Rica to Central Park. Together with his assistant and various biologists, he watched, identified, and took photos of anything visible to the naked eye within that space. The resultant photos were published in a breathtaking book, *A World in One Cubic Foot: Portraits of Biodiversity* (Liittschwager, 2012).

Inspired by Liittchwager's work, Smithsonian scientists and education specialists initiated the Biocubes project with partners from both academia and industry, including the University of Maryland, National Geographic, the Great Nature project, and iNaturalist. Researchers from the Smithsonian and University of Maryland lead the development of CyberSEES and NSF proposal writing.

The Biocubes project is first located in the context of science education. It provides a practical way for educators to prompt students and the public to get hands-on experience of collecting and documenting biodiversity observation data so their knowledge and

awareness of biodiversity can be increased. At the same time, students and the public also contribute to knowledge about biodiversity by sharing their biodiversity observation data with researchers and other members of the public who may be interested in using it. iNaturalist is the platform adopted by CyberSEES for documenting and sharing Biocubes data publicly.

A major method of recruiting data contributors for the Biocubes project was holding training workshops. Several beta workshops were held in 2013 and 2014 where educators were trained to build biocubes and collect Biocubes data. Once trained, the educators should be able to implement the Biocubes project in their classes, and these educators and their students become Biocubes data contributors (CyberSEES project description, 2015).

### 3.5 Data collection and Analysis

Data collection for this dissertation is guided by the three sub questions as well as the theoretical models/frameworks introduced in Chapter 2. Data collection and analysis are conducted simultaneously so that the developing analysis of the results, together with the theoretical models/frameworks, guide subsequent data collection and analysis.

#### 3.5.1 Research procedure

Figure 3.1 shows the research procedure including a timeline for collecting and analyzing data from the two case studies. This dissertation collected and analyzed data for the EOL case since Spring 2014. Based on the initial data analysis results, CyberSEES was

chosen as the second case. The data collection and analysis of this case started in Spring 2015.

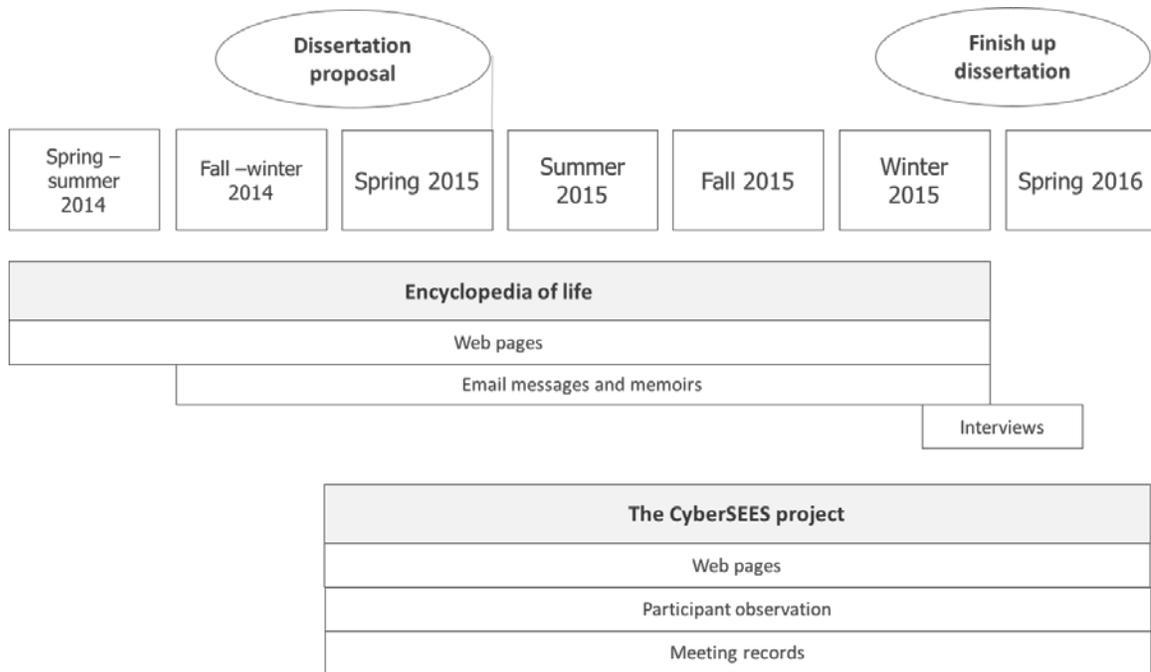


Figure 3.1 Research procedure and timeline for this doctoral research.

### 3.5.2 Data collection

Multiple sources of data were collected for each case. Data collection methods include (digital) artifacts, documentation, participant observation, and interviews.

#### *Artifacts*

Artifacts collected in this work are digital: they are publicly viewable web pages. These artifacts must contain the information needed to understand the data providers (i.e., EOL content partners), data mediators (i.e., human mediators, technology mediators), and the processes of data sharing by human actors. In the case of EOL, web pages (e.g., whole web pages, screenshots) and the information systems (i.e., data infrastructure, such as the

user account management system, data source management system) are collected. These web pages come from the EOL website, data providers' websites, and other kinds of websites, such as news websites and LinkedIn.

Similarly, in the case of CyberSEES, web pages and the information systems are collected. In addition, the database containing Biocubes data is also retrieved by using the data exporting function provided by iNaturalist.

### *Documentation*

Different types of documentation data are collected for both cases. In the case of EOL, this includes both public and private data. The publicly viewable data includes published academic papers, reports, and slides about EOL. Private data includes private communication messages between EOL staff, technicians, and content partners, and memoirs of EOL staff who developed the system and coordinated the communication between EOL and content partners. The communication messages and memoirs are collected from the JIRA system.

The JIRA system is an agile project management tool that was originally designed and developed for software development teams (Jira : Project Management Software, 2016). Its major features include enabling the project members to plan, track, and organize the work of developing software (e.g., tasks, ideas, people's requests). EOL staff adopted the JIRA system to help manage the work of building and maintaining partnerships with all the content partners. In the JIRA system, EOL staff created one or multiple "ticket(s)"

for recording the work in progress of building the partnership with each content partner. There are two major types of tickets: collaboration tickets and data tickets. Collaboration tickets are usually created when EOL starts to work with a new (potential) content partner and are used to record all the efforts made by EOL and content partner staff to establish a partnership. Data tickets are usually created for solving specific technology issues related to data. After data tickets are created, they are usually linked to the corresponding collaboration tickets as sub-task tickets.

The communication within messages and memoirs in JIRA tickets are in the form of comments left by EOL staff, since only EOL staff have access to the system. Content partner staff usually do not have access, unless they are also EOL staff or send a request for and are granted special access, which is very rare. Other than email, which are copied into system comments, all communication messages and memoirs are directly left in the comments by EOL staff.

Since it is impossible to collect and analyze all JIRA tickets for all EOL content partners, one method of randomly choosing the content partners for analysis of their JIRA system content was created and was implemented in summer 2014. Content partners were divided into seven groups based on the number of years they had partnered with EOL (less than 1 year, 1 year, 2 years, 3 years, 4 years, 5 years, and 6 years). Within each group, a random order was created for the content partners. The first 30% (N=78) of the content partners in each group were chosen as candidates whose JIRA system content might be collected later as documentation data. This sampling plan ensured that the

sample included content partners of differing length relationships with EOL. EOL had a total of 257 content partners in summer 2014.

In the end, 36 of the potential 78 content partners' JIRA system tickets were collected and analyzed to reach data saturation. The total number of JIRA system tickets that were analyzed was 111, containing a total of 937 analyzed comments and 647 analyzed emails.

In the case of CyberSEES, documentation data includes the project description, as well as documentations created for introducing the Biocubes project and guiding data collection (e.g., Biocubes data collection protocols, observation sheet).

#### *Participant observation*

Participant observation is only applicable to CyberSEES. The Biocubes project data collection training workshop held in Florida on January 24–26, 2015 was observed. I participated in and observed the whole workshop. When observing, field notes and photographs were produced, and partially expanded after the observation session ends. In addition, the biweekly meetings among CyberSEES project organizers (who are also CyberSEES project members) were observed since June 2015. Each meeting usually lasted about one hour. Meeting notes are recorded each meeting.

#### *Interviews*

Semi-structured interviews were used only for the EOL case. By using a purposive sampling method (Merriam, 2014), two core EOL staff who were responsible for

managing the data sharing relationships with EOL content partners were recruited as interviewees. The aim of these interviews was verifying and refining the findings from the analysis of the data from other data sources (i.e., artifacts, documentation). The interview times for the first and second interviews were 90 and 60 minutes. I conducted the interviews which were audio recorded, then partially transcribed.

For each interview, the interview started by introducing the findings from answering the three research questions with respect to EOL (i.e., who are the data providers, who are the data mediators, and what are the sharing processes). The summary and important details of the findings were presented as slides (see Appendix A) to the interviewees.

The interviewees were told to feel free to interrupt the interviewer if there was anything in the findings that did not make sense, was incorrect or misunderstood, or they wanted to comment or add more information. After the first interview, the findings were refined and this updated version was used in the second interview. After the second interview, the findings were refined again until both of the interviewees agreed to the contents in the refined findings.

#### *Data storage*

All hard-copy versions of the data collected were and continue to be stored in the authors' locked filing cabinet. Digital data was and is kept securely in different file folders on the author's personal computer. Data that are not public viewable and contain personally identifiable information were and are protected by password.

### 3.5.3 Data analysis

This dissertation offers a form of data triangulation by analyzing multiple data sources (Preece, Sharp, & Rogers, 2015). The theoretical frameworks and models (Figure 2.4, Figure 2.5, Figure 2.7, Figure 2.9, Figure 2.6, and Figure 2.8) discussed in Chapter 2 were used to guide initial data analysis. The initial coding schema developed from the theoretical frameworks can be found in Appendix B.

For the first and second research sub questions (“who are the data providers” and “who are the data mediators”), the model of identification (Figure 2.9) is used to guide the initial coding of the identification of data providers and data sharing mediators. For the third research sub question (“what are the data sharing processes”), the framework of data sharing processes (Figure 2.6) is used as initial coding scheme to identify different components of data sharing in each case. The framework of data sharing contexts (Figure 2.8) is used as the coding scheme for identifying the complex interrelationships among the human actors, the institutions/organizations/communities they belong to, and their data sharing culture and institutional logics. This framework also guided the initial coding of the social interactions among human actors that make data sharing possible.

The data were iteratively coded by first open coding using the methods introduced in Miles et al. (2013, p. 74–83) (i.e., elemental method, affective method, literacy and language method, exploratory method, and procedural method) to make sense of the data as a whole (Elo & Kyngas, 2008). Microsoft Word and pen and paper were used to conduct the open coding.

The data were then coded by using deductive content analysis (Kyngas & Vanhanen 1999; Elo & Kyngas, 2008) based on the theoretical frameworks and model mentioned earlier, moving from the general to the specific (Burns & Grove, 2005; Elo & Kyngas, 2008). Then inductive content analysis was used to code both the original data and deductive coding results (Elo & Kyngas, 2008), with a new and broader set of concepts emerging. In addition, to answer the third sub question specifically, deductive process analysis (Crowston, 2000) was also adopted. TAMSAalyzer was used as the coding tool for conducting deductive, inductive, and deductive process analyses (Weinstein, 2012).

In addition, for analyzing EOL JIRA system content, in addition to deductive and inductive coding, a new method was created and adopted by the author: using a color pencil on hard copy notebook to draw the interaction flows between the human actors from EOL and the content partners.

Table 3.2 summarizes the data sources and analysis methods adopted to answer the three research questions for each case.

Research questions	Cases and data		Data analyses
	EOL	CyberSEES	
<b>Q1: Who are the data provider?</b>	Webpages <ul style="list-style-type: none"> <li>• Content partners' own websites</li> <li>• EOL website</li> <li>• Other websites</li> </ul>	Observation data <ul style="list-style-type: none"> <li>• Field notes (Florida workshop)</li> <li>• Photos (Florida workshop)</li> </ul> Webpages <ul style="list-style-type: none"> <li>• iNaturalist website</li> <li>• Biocubes project page</li> <li>• Other websites</li> </ul>	<ul style="list-style-type: none"> <li>• Open coding</li> <li>• Deductive coding</li> <li>• Inductive coding</li> </ul>
<b>Q2: Who are the data mediators?</b>	Webpages <ul style="list-style-type: none"> <li>• EOL website</li> <li>• EOL working group members' organizations websites</li> <li>• LinkedIn website</li> <li>• Personal websites</li> <li>• Other websites</li> </ul> EOL JIRA system contents <ul style="list-style-type: none"> <li>• Emails</li> <li>• Internal messages (i.e., comments)</li> <li>• Memoirs</li> </ul>	Observation data <ul style="list-style-type: none"> <li>• Field notes (Florida workshop, biweekly meetings),</li> <li>• Photos (Florida workshop, biweekly meetings)</li> </ul> Webpages <ul style="list-style-type: none"> <li>• iNaturalist website</li> <li>• Biocubes project page</li> <li>• Other websites</li> </ul>	<ul style="list-style-type: none"> <li>• Open coding</li> <li>• Deductive coding</li> <li>• Inductive coding</li> </ul>
<b>Q3: What are the data sharing processes</b>	Webpages EOL JIRA system contents <ul style="list-style-type: none"> <li>• Emails</li> <li>• Internal messages (i.e., comments)</li> </ul> Interviews	Observation data <ul style="list-style-type: none"> <li>• Field notes (Florida workshop, biweekly meetings),</li> <li>• Photos (Florida workshop, biweekly meetings)</li> </ul> Webpages <ul style="list-style-type: none"> <li>• iNaturalist website</li> <li>• Biocubes project page</li> <li>• EOL website</li> </ul> Database exported from iNaturalist	<ul style="list-style-type: none"> <li>• Open coding</li> <li>• Deductive coding</li> <li>• Inductive coding</li> <li>• Drawing relationship map</li> </ul>

Table 3.2 Summary of multiple data sources and data analysis methods regarding answering the research questions for each case.

### 3.6 Data validation

This dissertation establishes the quality of research design from four aspects: construct validity, internal validity, external validity, and reliability (Yin, 2013, p. 45). Construct validity concerns whether the research design has the correct operational measures for the concepts being studied (Yin, 2013). This dissertation investigates the data sharing practices across research and public communities. The change in data sharing observed from the literature to real life is that data have become more accessible to a broader audience than before. The specific measure of this change can be the communities of data users. The data are shared with users beyond the research community, that is, anyone with the Internet access. The public community, defined in Soranno et al.'s (2015) round table model, is the community of data users focused on in this research. As Yin (2013) suggests, in order to increase construct validity of case study research, three tactics are adopted in this research: use multiple sources of evidence, establish a chain of evidence, and ask key informants to help review the draft case study.

Internal validity is important in this research because the research design includes investigating the influence of data sharing contexts (i.e., embeddedness) on the data sharing processes. As Yin (2013) suggests, pre-theoretical and analytical frameworks have been established (i.e., the frameworks of data sharing processes and data sharing contexts) for guiding the initial data analysis and increasing internal validity.

External validity deals with the concern of generalizability of case studies. The case study approach focuses on in-depth investigations of a limited number of cases, therefore

the results of this research cannot be considered representative and generalizable. Instead, as Yin (2013) suggests, replication logic is adopted and the two cases were carefully chosen to fill the theoretical categories in order to strive for external validity.

The goal of reliability in a case study method is to minimize the errors and biases. Although it is impossible for a researcher to be human error free and for a qualitative researcher be purely objective, some strategies are helpful to improve the reliability of case study research, or at least make it more transparent. Yin (2013) suggests creating a case study protocol and case study database to document the research procedures. By adopting these two tactics, this research ensured that its key steps are as operational as possible.

Furthermore, in order to provide more information to help judge the quality of this research, a short discussion about reflexivity (Bailey, 2007) is included here to illustrate how the author of this dissertation thinks her status characteristics, values, history, and decisions may affect the research results.

I am an international student who comes from China and moved to US four and half years ago to pursue my PhD degree. My interest in biodiversity had been developing since I was a child, long before coming to the US. The area of China I am from, Yunnan province, has the most diverse natural resources in China, especially birds and mammals, although Yunnan is only 4.1% of China's total area (Yang, Tian, Hao, Pei, & Yang, 2004). I was brought up with pride and love for the biodiversity of my province. This is

where my personal interest and care for biodiversity and the natural environment originally comes from and it only I grew as I got older.

When I realized how badly the biodiversity in the world suffered from human beings' brutal exploration and damage (as introduced in the earlier section, domain selection), I decided to devote my research about data sharing to the domain of biodiversity. I believe that to a large degree, human beings' indifference to the biodiversity and the natural environment is due to their limited access to biodiversity knowledge and restricted real-life experiences of getting close to nature. Therefore, I hold a positive view of the value of sharing research data with not only researchers but also non-professionals, and I uphold the value of citizen science. I believe sharing research data with non-professionals as well as including them in collecting data has significant benefits to not only advance scientific research, but also to increase the public's awareness of biodiversity and the nature environment in general. These facts about my history and values influence my motivation and attitude in doing my dissertation in a positive way.

In addition, I would also like to briefly discuss my position in the two cases. I did not participate in developing the knowledge infrastructures of either case. In the case of EOL, I was able to work with one of the EOL staff because of previous research opportunities not directly related to developing EOL human or technology infrastructures. At that time, EOL was already a mature and successful biodiversity repository. The positive side effect of the previous research opportunities was that they provided me a chance to learn how the human workers worked on developing and

maintaining a knowledge infrastructure—it was much harder than I could have imagined. This inspired my thinking that these human collaborative efforts were still far from being recognized and acknowledged well enough. It also made me think that it is no wonder that there are a very limited number of biodiversity repositories like EOL targeting not only researcher users but also non-professionals user in the public community. One possible reason for this could be because people know very little about what how to build such a knowledge infrastructure.

For the second case, after I got preliminary results about the collective-level data providers in EOL, I was looking for another case that could allow me to study individual-level data sharing. Coincidentally, near that time, I got an opportunity to join a new research project (i.e., the CyberSEES project) as a research assistant to help study infrastructure development and design for sharing citizen science data. Furthermore, my role in the Biocubes project training workshop in Florida was as one of the workshop facilitators.

Enabling Biocubes data to be shared on the EOL repository is not CyberSEES's major intention or goal, but more a convenient condition because EOL and iNaturalist had already built the partnership long before CyberSEES was launched. Although I noticed this “unintentionally,” the important alignment makes the perfect example to show how data created by non-professionals could travel to an authoritative aggregator repository with many other data created by researchers. More importantly, every step of sharing is for both researchers and non-professionals. Of course, this travel could not happen

without large amount of effort made by human mediators who make up human infrastructure.

This history about my positions in both cases helps to understand the degree of my “bias” in choosing the cases. I did not intentionally not choose any other specific cases.

Last but not least, as a foreign researcher who was raised in Chinese culture and whose first language is not English, I set out to investigate research questions which can be answered as objectively as possible (i.e., social identities, processes), and were less likely to be influenced by my personal cultural background. My background made me more sensitive about getting confirmation of my data analysis and findings from my advisor and the core informatics in the two cases, as well as making sure that I understood my data correctly.

### 3.7 Conclusion

The goal of this chapter was to explain the research design and methods adopted in this dissertation. This chapter first explained why answering the overarching research question (How data are shared effectively across research and public communities?) needs to answer the three sub research questions (Who are the data providers? Who are the data mediators? And what are the data sharing practices?). The relationships among the three sub research questions were clarified as the three essential parts of data sharing practices. Data sharing practices that make the data sharing occur across research and public communities is the answer of the overarching research question.

This chapter went on to explain why this dissertation:

- chose to use the case study method;
- chose to focus on the biodiversity domain;
- choose the Encyclopedia of Life and the CyberSEES Project as the two real-world cases.

At the end, this chapter introduced the details of the multiple data sources that were collected for this dissertation and the methods and tools that were used to analyze the data and ensure the research quality.

## 4 Case one – Encyclopedia of Life

### 4.1 Overview of the findings

EOL is an open access online database (Parr et al., 2014) and a content curation community (Rotman et al., 2012). This dissertation introduced EOL as a large-scale aggregator repository in Chapter 3. The three terms database, community, and repository are used to refer the products of the EOL knowledge infrastructure. The EOL website (eol.org) hosts and makes these products accessible to data users and, while they are part of the knowledge infrastructure, they cannot represent the whole of it. More specifically, these products are the important parts of the technology infrastructure within the entire EOL knowledge infrastructure and it is through them that biodiversity data are shared with users across research and public communities. The findings in this chapter demonstrate how data are shared through these products by answering the three sub questions: who are the data providers, who are the data sharing mediators, and what are the sharing processes. The invisible part of EOL knowledge infrastructure, the human infrastructure, is revealed through answering these questions.

EOL as a whole entity has a clear vision to provide “global access to knowledge about life on Earth” (What is EOL? - Encyclopedia of Life, n.d.). Their mission is “to increase awareness and understanding of living nature through an Encyclopedia of Life that gathers, generates, and shares biodiversity knowledge in an open, free accessible and trusted digital resource” (What is EOL? - Encyclopedia of Life, n.d.). Every effort made by EOL staff contributes to fulfilling EOL’s vision and mission.

The knowledge about life on Earth, “biodiversity knowledge,” displayed on EOL’s website is comprised of the taxon information shared by data providers. These providers are EOL’s “content partners” and are described as follows:

*“Encyclopedia of Life content partners have large amounts of information about biodiversity in their own websites or databases that they also share via EOL pages. We (EOL) greatly appreciate their critical contributions to the EOL mission”* (Content Partners, n.d.)

Content partners’ data is shared via EOL after building a formal collaborative sharing partnership. Human actors (i.e., staff) from both EOL and the content partner are first connected, after which the processes of building the partnerships, transferring the data to EOL, and displaying the data on EOL’s web pages begins. By March 2016, EOL has 329 content partners. The number of the content partners continues to grow.

The three questions were asked to investigate how data providers shared their data with EOL:

- Who are the data providers (i.e., EOL content partners)?

This dissertation collected and analyzed textual content about content partners from their own websites and databases. Similar textual materials on EOL’s and other websites (e.g., news websites) were also collected and analyzed. These websites and databases are publicly viewable. Analysis indicated who they are,

what they do, and what they value. The answers to this question reveal the data sharing contexts within which data shared with EOL were originally located.

- Who are the data sharing mediators?

Data sharing mediators connect data creators and users by making creators' data available to users. Data sharing mediators are both human and technology mediators; therefore, answering this question comprises two parts: who are the human mediators and what are the technology mediators?

For answering who the human mediators are, the textual content of EOL JIRA tickets for the 36 selected content partners were collected and analyzed. These tickets are private documentations that only EOL staff have access to. JIRA is an agile project management tool is designed and developed for software development teams (Jira : Project Management Software, 2016) so EOL adopted it to manage the work of building and maintaining data sharing partnerships with the content partners. The results of analysis reveal the organizational structure of EOL human mediators and the organizational identities of both mediators from EOL and the content partners.

To answer what technology mediators are, the information systems (i.e., data infrastructure, such as the user account management system and data source management system) of EOL's knowledge infrastructure are explored and

analyzed. The answers to this question reveal both visible and invisible parts of the data mediators that enable data sharing from the content partners.

- What are the data sharing processes?

Data used for the second question—textual contents in EOL JIRA tickets for the 36 content partners—are reanalyzed for different purposes and by different methods. This analysis reveals the details of the efforts made by the data providers (i.e., EOL content partners) and data sharing mediators.

The remainder of this section will report the details of the answers to each question.

## 4.2 Answering the first question: who are the data providers?

### 4.2.1 Diversity of data providers

In a pilot study, six non-mutually exclusive types of collective-level data providers were identified: venerable organizations, professional repositories, citizen science initiatives, social media platforms, education communities, and subsidiaries (He et al., 2015).

Although one data provider could show characteristics of two or more types, their most notable characteristics and primary focus are used to categorize them into one type.

The 275 content partners' web pages that were analyzed in the pilot study were revisited to confirm the six types of content partners discovered in the pilot study. This reanalysis confirmed that the understanding of the content partners' identities in this dissertation is consistent with that of the pilot study. By the time of dissertation analysis in March

2016, EOL had 329 content partners. Using the same analysis method, the website contents of the 54 new partners were analyzed, completing the analyses of all 329 content partners. Among the 54 new partners, a new type was identified: academic papers. Therefore, there are a total of seven types of content partner (i.e. data provider).

#### *Venerable organization*

Venerable organization data providers are traditional and authoritative professional organizations that have strong reputations and long histories (e.g., from 70 years to more than 260 years). These organizations' identities are rooted deeply in an offline space. Benefit for mankind and societal good are highly valued, and their behavior is complex but tightly tied to these values. Typical venerable organizations include environmental organizations, natural history museums, and government entities.

#### *Professional repository*

Professional repository data providers are professional databases or repositories. They usually have a large collection of data, including one or multiple databases. This type of data provider proliferated from the 1990s onwards. Most of these repositories are initiated, hosted, or supported by organizations, universities, government agencies, or institutionally situated research teams. Their identities emphasize the content (i.e., data) they want to share in online environments rather than their affiliation, reflecting their primary goal and mission. Some of these data providers are professional databases that focus on sharing data only, while others are communities of practice with simple social features that encourage professionals to share their data, communicate, and collaborate.

It is worth mentioning that although most professional repository data providers adopt different platforms, many share their data via one of two platforms: Lifedesk (lifedesks.org) and Scratchpads (scratchpads.eu). There are over 80 professional repository data providers using Lifedesk and Scratchpads among the 329 content partners. This number continues to increase.

Lifedesk and Scratchpads platforms are designed and developed for providing researchers online space and tools to manage and share biodiversity research data publicly. These platforms not only enable researchers to share data by displaying it on web pages, but also support sharing data by easily creating the data source files, which are needed when researchers want to share the same data on other platforms. These two platforms are easy to use without requiring users to have a technology background.

#### *Citizen science initiative*

Citizen science initiative data providers gather data from non-professionals for scientific use. This type of data provider usually has two parallel goals: encouraging members of the public to contribute data that can be used by researchers, and meeting other people who are interested in nature. These data providers operate primarily online, may not be affiliated with other organizations or research projects, and can function as communities of interest. They usually do not require data contributors to provide an authentic identity; therefore, while it might be evident how the data creators provide data, who they are might not be.

#### *Social media platform*

Social media platform data providers, such as Flickr, YouTube, and Wikipedia, are popular social media platforms that appeared in the 2000s. They encourage a broad range of users to share different types of information for diverse reasons; they were not originally created for scientific use. The data creators from this kind of data provider might not be aware that their data have research value, or have a potential to become research-grade data. However, it is still possible to use social media platform data providers as placeholders of data which might reach research-grade and therefore be relevant to EOL. Just like citizen science data providers, social media platform data contributors' authentic identities may be unknown.

#### *Education community*

Education community data providers include classes run by education institutions, groups of students, and data collections contributed by students in the context of science education. Educators adopt online communities as educational data management tools to assist the students' learning process and improve their scientific learning motivation and outcomes.

#### *Subsidiary*

Subsidiary data providers are a special type of provider. Trained biologists, some of whom worked for EOL, created these data repositories, the primary reason for which was to increase the comprehensiveness of the data on EOL. Although some data providers are semi-autonomous, each of them is a fully identified unit distinct from other subsidiary data providers.

### *Academic paper*

Among the 54 new content partners, a new type was identified: academic paper. In the EOL profile web pages for this type of content partner, they are introduced merely as citations of academic papers. These data providers share research data through academic papers, which might have been shared in professional repositories before being shared with EOL.

### *Other*

A few data providers could not be categorized into one of the seven types. Most of these did not have their own website or database and may only have limited introduction information on the EOL website. Therefore, there was not enough information to confidently recognize who they were and to which type of data provider they should belong. For those data providers who have their own websites but could still not be categorized, it is likely because their identities are less common, such as a journal, research project, tool, or personal photo collection.

Figure 4.1 shows the number of data providers that are categorized in these seven types. The professional repository data providers (N=148) and subsidiary data providers (N=95) represent the majority. They account for 44.98% and 28.88% of total data providers respectively. The numbers of venerable organizations (N=20), education communities (N=13), and academic papers (N=20) account for 6.08%, 3.95%, and 6.08% of total number of data providers respectively and therefore account for a similar proportion of

providers. There are a limited number of citizen science initiative data providers (N=5) and social media platform data providers (N=7) and, as such, only account for 1.52% and 2.13% of the total number of providers respectively.

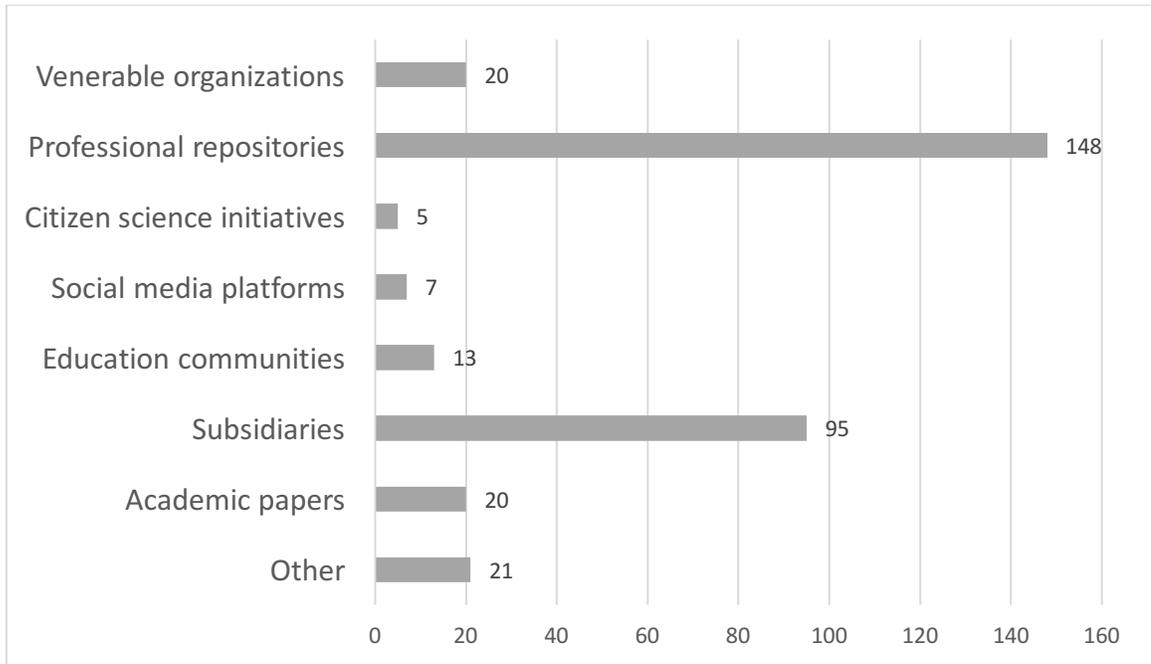


Figure 4.1 The number of different types of data providers. The total number of data provider is 329.

The results of the analysis of the website contents show a diversity of data providers. The seven types represent different data sharing contexts where data was located before they were shared on the EOL website. This diversity also indicates that there are different cultures and institutional logics of data sharing that underlie these data providers.

The website content did not provide enough information about who the key human actors are that enable data sharing and by what kind of effort. These key human actors are the mediators who transfer data from the data providers' own databases to the EOL website.

Therefore, by adding a new data resource (i.e., EOL JIRA system content) into the analyses, the key human actors will be revealed.

### 4.3 Answering the second question: who are the data sharing mediators?

“Data sharing mediator” is a sociotechnical concept. Data sharing mediators in a knowledge infrastructure include two parts: human mediators and technology mediators. The relationship between these two is that human mediators collaboratively use technology mediators to make the data creators’ data available to users in an online environment.

#### 4.3.1 Human mediators

In generally, all human actors who work to make creators’ data available to users can be considered a human mediator. However, this dissertation focuses on the human mediators who work directly with the content (i.e., data) shared on the platform. They are the key human actors who ensure data sharing happens. The human mediators exist in both EOL and the data providers (i.e., content partners).

##### 4.3.1.1 Human mediators from EOL

The governance model of the EOL knowledge infrastructure comprises individuals and organizations who share the same belief and desire of promoting large scale biodiversity knowledge sharing in an open, freely accessible, and trustworthy online environment (EOL Governance, n.d.). From the perspective of institutional logics (Thornton et al., 2012), the organizational identities of human mediators are mainly shaped by their

affiliate organization and EOL. Identities include who they are, what they do, and their focus of attention (i.e., what is important for them, and what they value) (Ashforth et al., 2008). Therefore, it is important to understand EOL organizational structure before we identify specific human mediators.

The governance model of EOL includes three major parties rooted in different organizations that operate and manage EOL knowledge infrastructure: 1) the EOL Executive Committee; 2) the EOL Secretariat; and 3) EOL Working Groups (EOL Governance, n.d.). The members of the EOL Executive Committee are EOL's Board of Directors and are a group of senior figures from EOL's cornerstone institutions, regional, national, and thematic EOLs, and other major financial or in-kind partners. Cornerstone institutions are the leading scientific organizations around the world that contribute significantly to biodiversity research and conservation. The Committee members focus on providing high level governance in terms of the long-term sustainability and success of EOL knowledge infrastructure.

The EOL Secretariat and Working Groups are responsible for day-to-day operation and management work. The Secretariat includes a project manager (i.e., director of operation), project coordinators, and the administrators of EOL. They directly report to the EOL Executive Committee and coordinate and plan the Working Groups, whose daily work concentrates on managing and delivering EOL components.

Of these three parties, the Working Groups directly work on setting up an online environment and connecting data creators and data users in EOL's online environment. These groups, of which there are three, have a more complex structure than the previous two parties, given that the members of these groups are based at different organizations and represent three major focuses of EOL's day-to-day operation and management work (EOL Governance, n.d.). EOL's three Working Groups are:

- the Biodiversity Informatics Working Group (BIG) (Biodiversity Informatics Working Group, n.d.);
- the Species Page Working Group (SPG) (Species Page Working Group, n.d.);
- the Learning and Education Working group (LEG) (Learning and Education Working group, n.d.).

Most BIG members were developers from the New Library of Alexandria and the Marine Biological Laboratory in Woods Hole. They developed and administrated EOL technology infrastructure with the help provided by contributing developers around the world (e.g., contractor developers and developers from the content partners). The infrastructure they developed allows human mediators to gather the data shared by hundreds of content partners, organize these data, and present them on the EOL website.

SPG is a group of biologists led by a core based at the Smithsonian Institution's National Museum of Natural History. The core group, who call themselves SPGers, includes the Director of SPG and two species page coordinators. SPGers can identify and report

trustworthy and valuable biodiversity content. The core group then works on behalf of EOL to build partnerships with the owners/managers of this content to share trustworthy data. After the partnerships are successfully built, these contents (i.e., data) can be transferred to species pages on the EOL website. SPGers are the key human actors who ensure an open access environment by managing EOL's intellectual property rules.

SPGers are also responsible for building and maintaining the EOL curators' community. Curators, invited by SPGers to improve the quality of data shared by the content partners, could be biologists or experienced citizen scientists. EOL curators manually check the contents shared on the EOL website, and report any errors to the SPG. They also help identify good quality data still marked as "unreviewed" and promote them to trusted status.

LEG members are educators and scientists from Harvard University and the Museum of Comparative Zoology at Harvard. They do not work on connecting the data providers to EOL but instead focus on connecting data users. They explore and promote worldwide educational uses of EOL two ways: 1) seeking opportunities for EOL to serve educators, citizen scientists, and students; and 2) encouraging the development of new tools and apps that facilitate biodiversity information sharing.

In addition to these three groups, there are two additional Working Groups: the Scanning and Digitalization Group (directed by the Biodiversity Heritage Library) and the Biodiversity Synthesis Group (Blaustein, 2009). However, at the time of this analysis,

they were not listed as major working groups on the EOL website. They were also not listed on EOL's government model documentation web pages as core Working Groups. Therefore, these two groups are not considered in this dissertation.

The descriptions of EOL's organizational structure were based on EOL website content, showing how and in what ways EOL and these Working Groups would like to introduce themselves to all online data users. They represent the umbrella groups of human mediators who create EOL knowledge infrastructure. In order to gain a deeper understanding of individual data mediators' work, the EOL JIRA system contents were added as a new data source.

In the EOL JIRA system, members from BIG and SPG appeared as the active users. They created JIRA tickets, left comments in the tickets, and added attachments to the tickets. The core group members from SPG (i.e., the Director and two species page coordinators) were the most frequent JIRA users. Members from LEG were most infrequent JIRA users, rarely creating tickets and leaving very limited number of comments. Because JIRA was adopted to support partnership management and connecting data providers to EOL, LEG had little role to play given their focus of attention to connect data users in educational contexts to EOL.

Although LEG made significant contribution to EOL by identifying various precious opportunities to use the data shared on EOL for education purposes, this dissertation focuses on investigating the human mediators who directly work on making the

connection between the data providers and EOL. In a future study, the human mediators from LEG and their work facilitating the use of EOL in the education community will be considered.

Therefore, this dissertation focused on studying the core human mediators from BIG and SPG since they are the key human actors who ensure data sharing actually occurs.

BIGers and core SPGers work closely to present the data on EOL species page. Both groups have liaisons and project coordinators—also the human mediators—to help smooth the collaboration between the two groups.

In addition, JIRA system content revealed that there are other human mediators who worked under the leadership and guidance of BIG and/or SPG. They directly participated in the work of facilitating data sharing, but do not have explicit affiliation relationships with any of the three Working Groups. These human mediators are EOL fellows and contractor workers.

EOL fellows are biologists enrolled in the EOL Rubenstein Fellows Program (EOL Rubenstein Fellows Program, n.d.). EOL launched this program for two major reasons: 1) support the development of targeted content, mainly contributed by professionals, about particular groups of organisms; and 2) enlarging the impact of original biodiversity research by scientists at the early stage of their careers (e.g., postgraduates, graduate students). This program provided funds to support fellows to transform research related databases and media elements into rich online resources that can be shared on EOL.

Fellows are also supported and encouraged to engage in collaborations with a wider range of colleagues in research community.

The then-current EOL fellows worked closely with SPGers. By following the leadership of the core members, the fellows contributed biodiversity content directly to the species pages on the EOL website. They represent a large number of the subsidiary data providers (one of the seven types of collective-level EOL content partners).

Developing and managing EOL technology infrastructure and the large amount of content shared by diverse data providers require huge amount of effort by human actors. Relying on the members of the Working Groups alone was not enough. Therefore, EOL hired contractors to help with their day-to-day work. Contractors include technicians who focus on technological tasks (e.g., linking the content partners' data to the EOL website) and biologists who help with biodiversity content (e.g., sorting and organizing the biodiversity data on species pages on the EOL website). Contractors work closely with both BIG and SPG members.

Figure 4.2 shows the organizational structure of the human mediators within EOL. This figure also shows the relationships between the Working Groups, the organizations of each group, and the individual human mediators. The organization information is in parentheses. The individual human mediators are in dashed line boxes.

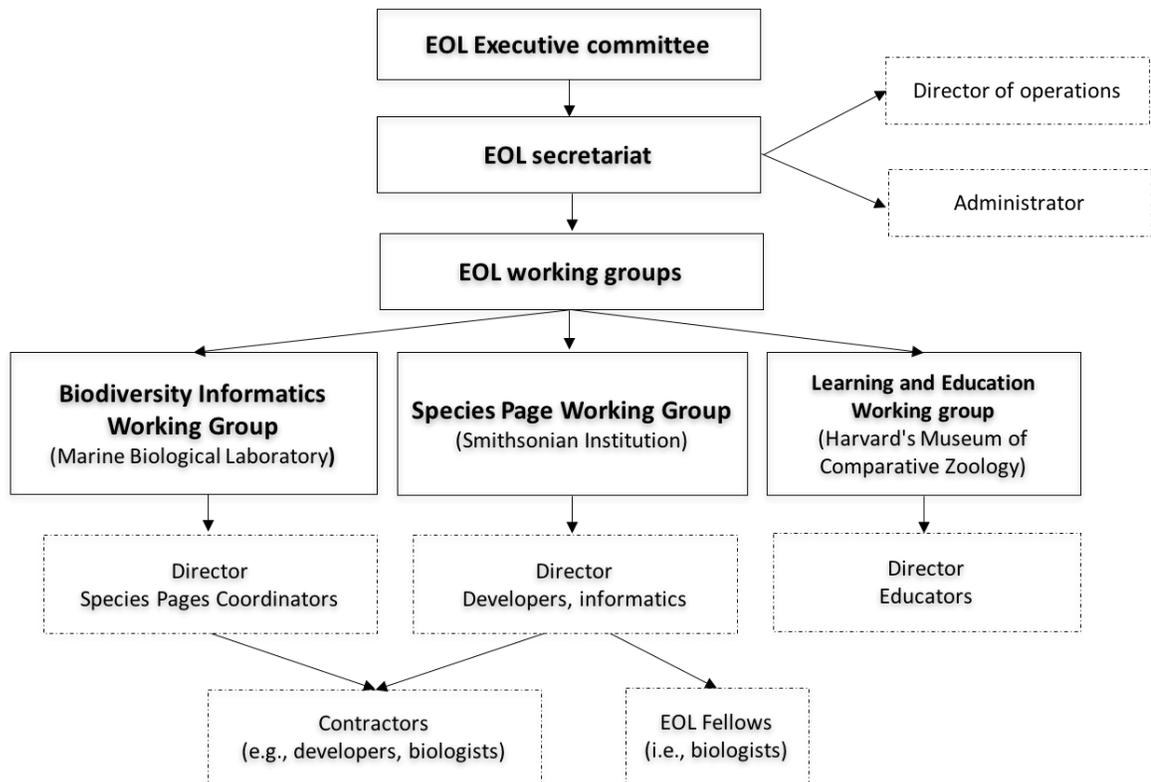


Figure 4.2 EOL organizational chart

#### 4.3.1.2 Mediators from content partners

The analyses of the EOL website content and the JIRA system revealed the full picture of organizational structure within which the human mediators of EOL are embedded. This structure indicates that the EOL knowledge infrastructure was created based on collaboration among different organizations. However, for content partners, only a small piece of their own organizational structure was revealed because the human mediators from content partners are only able to reflect a small part of the organizational structure.

However, by analyzing the content on the EOL website, content partners' own websites, other websites (e.g., LinkedIn), and the JIRA system, it was possible to identify the

following information about the human mediators from content partners: who these individual persons were, which organization or institution they worked for, what they were capable of doing and willing to do to share data with EOL, and what was important to them in terms of sharing the data they managed or owned.

The organizational identities of individual human actors from the content partners who worked with EOL on collaborative data sharing were identified and categorized into three major types:

- Data managers (i.e., administrators);
- Technicians (i.e., developers);
- Data contributors (i.e., data creators/authors/editors/owners).

Data managers are administrative-level human actors from content partners and could be directors of organizations or institutions, leaders of communities, or a group of people of any size. They have the power to make decisions on behalf of the content partner and could be its sole representative. They are responsible for having high level conversation with EOL human mediators about building a data sharing partnership that matches and contributes to the content partners' goals and missions. They have control of assigning other human actors from the content partner to work on building and maintaining the partnership and introducing them to EOL human mediators.

Technicians are executive-level human actors from the content partners. They are also known as programmers or developers and are usually led by data managers. They are

responsible for developing and maintaining the repository, platform, and/or tools for storing and sharing the content partners' data. They represent the IT intelligence of the content partners.

Data contributors are any individual human actor who can make a direct contribution to the contents/elements contained in the data, including data creators, authors, and editors. Data creators are those who go to the field to collect biodiversity data in any form (e.g., physical, digital, or both). Data authors are human actors who might not collect field data, but instead digitize physical data, write descriptions, annotate, or provide other first-hand useful information to introduce and explain the data. They make the data understandable to others who might want to view and use the data. Data editors are human actors who, again, might not collect field data, might not write the first description of the data, but instead refine the data or revise the existing description to meet the needs of different users with different knowledge backgrounds and levels.

Each of the three types of data contributor could be the human actor who uploads the data to the online environment. All the human actors might have ownership of the data and the control of the data to some degrees. However, the human actors who upload the data do not necessarily have to be the data contributors themselves: data managers and technicians who do not directly contribute to the contents of the data can be also upload data.

#### 4.3.2 Technological mediators

The technological mediators discussed in this dissertation are focused on providing access to the data from diverse data providers to a wide range of audiences across research and public communities. Technological mediators are responsible for: 1) aggregating and integrating the data provided by diverse content partners in EOL database; 2) exhibiting the data in a way that everyone, regardless of biodiversity knowledge, can understand what it means; and 3) allowing users to export the data. This dissertation explores the entire EOL technology infrastructure and identified that technological mediators include three major information systems embedded in the EOL infrastructure:

- the Content Partner Management System;
- the EOL species page;
- the TraitBank system.

#### 4.3.2.1 The Content Partner Management System

The content partner management system is built for aggregating data from diverse data providers (i.e., content partners) into the EOL database. In this system, a data sharing partnership can be initiated by a regular EOL user who has registered on the website and has a member account. After an EOL member logs in, there is a Content Partner tab in his or her profile page where, upon clicking, s/he will see a new page with a button called “add new content partner” (Figure 4.3). By clicking this button, the member will be led to another web page that requires the EOL member to provide basic information about the data sources s/he would like to share with EOL, such as “Project name,” “Project

description,” and “Description of data.” After providing this information, the member clicks “create content partner” to create a content partner account.

A content partner account is different from a regular member account, but remains embedded in the regular account. Each content partner has its own content partner account. The content partner’s data source(s) is be uploaded to this account and then displayed on the EOL website. The whole process (i.e., from making the decision to set up this content partner account to formally publishing the data on the EOL platform) involve a significant amount of collaboration and cooperation by human mediators within EOL and between EOL and content partners. The following sections sets out important details of EOL data sharing processes: in other words, how human mediators use technology mediators to build a partnership and subsequently transfer data from the content partners to data users.

## Content Partners

Add a new content partner

### Content partner profile information

Content partner details managed by



[yrhe](#)

• **Project name**

**Project abbreviation**

**Display name** (optional)

Provide an alternative publicly visible name to be used instead of your *Project name*.

**Project URL**

• **Project description**

**Description of data** (private)

Please provide us with a description of the type of information you have, how many items are available, what organisms they relate to, and whether they have been assembled or checked by qualified experts or by a knowledgeable community.

**Project notes** (private)

**Logo**

Current logo:



Upload a new logo:

No file chosen

[Cancel](#)

Figure 4.3 The content partner account screenshot.

#### 4.3.2.2 The EOL Species Page

The species pages, or taxon pages (The Taxon Page, n.d.), on the EOL website are the centerpiece of the knowledge infrastructure (Blaustein, 2009). The idea of creating these species pages came from the biologist E. O. Wilson from Harvard University who had a vision of a web-based encyclopedia for all species in which each species page would “summarize everything known about the species, from its genome and proteome to its distribution, habitat, and ecological relationship, as well as ‘its practical importance for humanity’” (Blaustein, 2009, p. 551). At the time of this dissertation (Fall 2016), EOL contains more than 1,346,000 species pages.

Although these webpages are called “Species Pages,” it does not mean that each is limited to the species level. These page range from kingdom to the species level. For example, there is a species page for Bird (scientific name, *Aves*) at a class level (Aves, n.d.), and a species page for the Emerald-chinned Hummingbird (*Abeillia abeillei*) at a species level (*Abeillia abeillei*, n.d.). Each species page displays biodiversity information that is gathered from hundreds of diverse data providers (i.e., content partners). As the number of content partners grows, species pages keep evolving and updating so that they can convey additional and updated information to data users.

The interface of each species page structures different types of data carefully. Tabs for each different type of content (i.e., data) enable users to navigate through the species page (Figure 4.4). For example, the default “Overview” tab—the first to be displayed on arriving to a species page—acts as gateway page to help users quickly understand what

this organism is and looks like and what other information about it can be found on EOL. This page achieves this by providing “quick facts” and definitions of professional terms for those without specialist knowledge. Also provided is rich provenance information for the data, such as highly technical metadata (Parr et al., 2015), of interest to scientists, experts, and those in academia.

On the “Overview” tab, visual information (e.g., photographs and images) about an organism is emphasized and posted in the most conspicuous position on the page, as visual content is probably the most direct and simple way to communicate information about an organism to a user. Photographs and images easily convey information about the shape, color, and other visual features about an organism to most data users.

Besides the visual content, the overview page also contains other types of information about the organism, such as media data (e.g., images, video), trait data (e.g., geographic distribution, physical attributes, ecology and conservation data), scientific classification, maps, a comprehensive description, data resources, the community of people who are interested in this organism, and social interactions of EOL curators and regular users (e.g., curating activities, updating contents, comments). If a user is interested in learning more about the organism, they can click other tabs, based on their need (e.g., scientific, educational, interest), to access more information. Data providers’ information is clearly presented and credited on each species page so that users can easily track the origin of the content.

*Ailuropoda melanoleuca*  
Giant Panda learn more about names for this taxon

add to a collection

Overview Detail Data 170 Media 6 Maps Names Community Resources Literature Updates



*Ailuropoda melanoleuca* **THREATS**  
  © Smithsonian Wild  
 Source: [ESOL Images](#)

**IUCN threat status: Endangered (EN)**

**Comprehensive Description** [read full entry](#)

[learn more about this article](#)

**Description of *Ailuropoda melanoleuca***

*Ailuropoda melanoleuca* is the giant panda, a kind of bear that is native to central-western and south western China. The giant panda has a body shape typical of bears. It has black fur on its ears, eye patches, muzzle, legs, arms and shoulders. The rest of the animal's coat is white. Though it is classified among the Carnivora, its diet is mostly bamboo. Occasionally they eat other grasses, wild tubers, or even meat in the form of birds, rodents or carrion. It lives in a few mountain ranges in central China, mainly in Sichuan province, but also in the Shaanxi and Gansu provinces. Due to farming, deforestation and other development, the panda has been driven out of the lowland areas where it once lived. The panda is an endangered species, and needs active conservation measures. In 2007, an estimated 239 pandas lived in captivity inside China and another 27 outside the country. Wild populations probably number between 1500 and 3,000. Adults measure around 1.2 to 1.8 meters (4 to 6 ft) long, including a tail of about 13 cm (5.1 in), and are 60 to 90 centimeters (1 ft 10 in to 2 ft 10 in) tall at the shoulder. Males weigh up to 160 kilograms (350 lb). Females are 10&20% smaller than males. The average adult weight is 100 to 115 kilograms (220 to 250 lb). The giant panda has large molar teeth and strong jaw muscles for crushing tough bamboo. In addition to 5 fingers, the paw has a thumb modified from the sesamoid bone. The thumb helps the giant panda to hold bamboo while eating. The giant panda typically lives around 20 years in the wild and up to 30 years in captivity.

**THREATS**   David • Source: [BioPedia](#)

**Present in 91 collections** [see all](#)

- Species in China**  
1 other item
- Endangered Species**  
1 other item

**Belongs to 4 communities** [see all](#)

- EOL Biodiversity Informatics Group**  
17 other items, 12 members
- High Plain Elementary School**  
52 other items, 3 members
- Help The World**  
6 other items, 2 members

**EOL has data for 42 traits** [see all](#)

clutch/brood/litter size	(average) 1.5 1.62
body length (VT)	(average) 1,345.98 mm (adult)
home range	(average) 3.55 km² (average) 3.76 km²
body mass	(average) 104.30 g (newborn animal) (average) 117,000.00 g (adult) (average) 21,909.98 g (newborn)
behavioral circadian rhythm	nocturnal/crepuscular, cathemeral, crepuscular or diurnal/crepuscular
onset of fertility	(average) 2,152 days (female) 2,192 days (male) 2,413.02 days
precipitation in geographic range	(mean) 101.81 millimeters (per month)
population trend	Decreasing
habitat	broadleaf forest bome city mountain

**Classification**

Classification from IUCN Red List selected by [Cindy Parr](#) - [see more](#)

- [Archonta](#)
  - [Carnivora](#)
    - [Mammalia](#)
      - [Carnivora](#)
        - [Ursidae](#)
          - [Ailuropoda](#)
            - [Ailuropoda melanoleuca](#)

**Reviewed by 5 curators** [learn how to curate](#)

- [Michael Franke](#)
- [Katie Schulz](#)  
EOL content hunter & gatherer
- [Rob Mutch](#)
- [Deniz Martinez](#)  
#AmA Naturalist
- [Yan Wong](#)  
Evolutionary biologist

**Latest updates** [see all](#)

- [Yan Wong](#) changed the thumbnail image of "File:Xian Linyu in San Diego Zoo - Foto 4.jpg".  
15 DAYS AGO  
[reply](#)
- [Yan Wong](#) changed the thumbnail image of "File:Giant Panda in Beijing Zoo 1.JPG".  
15 DAYS AGO  
[reply](#)
- [Yan Wong](#) changed the thumbnail image of "File:BabyPandaASOZ.jpg".  
15 DAYS AGO  
[reply](#)
- [Yan Wong](#) changed the thumbnail image of "File:Chengdu pandas-d16.jpg".  
15 DAYS AGO  
[reply](#)
- [Deniz Martinez](#) marked "File:Giant Panda 2004-03-1.jpg" as trusted on the "[Ailuropoda melanoleuca \(David, 1869\)](#)" page.  
ABOUT 1 YEAR AGO  
[reply](#)

[post comment](#)

EOL content is automatically assembled from many different content providers. As a result, from time to time you may find pages on EOL that are confusing. To request an improvement, please leave a comment on the page. Thank you!

Figure 4.4 An example of EOL species page for Giant Panda screenshot (Overview page) (Giant Panda, n.d.).

#### 4.3.2.3 The TraitBank System

Having different types of data from diverse data providers aggregated, carefully organized, and then exhibited on a publicly viewable species pages is an effective way of sharing data with any potential data users with different knowledge backgrounds from different communities. Nevertheless, the EOL human mediators did not stop the sharing only at the exhibition level. They made further effort to enable the next level of sharing: allowing users to directly download the data.

When the EOL team designed and developed a data download system, they needed to determine what and how data should be available to users. Usually data users do not need to download all of the diverse data types presented on the species pages.

Furthermore, some types of data are either copyright protected or too large to be directly downloaded.

Users who want to download a batch of biodiversity data (e.g., a database) are much more likely to be researchers and experts than casual users. Therefore, the EOL team developed a data download system called TraitBank (Search TraitBank, n.d.) mainly based on the needs of data users from the research community. EOL human mediators chose to focus on sharing “trait” data because they realized that there is a strong need from the research community to receive aggregated, measurable characteristic data at the species and class level, for example for large scale modeling (Harfoot & Roberts, 2014; Parr et al., 2015). As a consequence, the first data download system to be embedded in the EOL technology infrastructure was designed and developed to enable data users to

download trait data provided by content partners.

The word “trait” refers to “any measurable characteristic, phenotype, property, or attribute of individuals or groups of the same taxon (type of organism)” (Parr et al., 2015, p. 1), such as mass, width, volume, and so on. EOL further emphasized that “at the heart of each trait record is an Occurrence, where the identity of the taxon and context in which the trait was observed or measured may be recorded (e.g. geospatial information, dates, life stages, individual counts).” (Parr et al., 2015, p. 3).

The TraitBank system allows data users, who must be registered EOL members, to download data through the download function initiated from the search facility (Parr et al., 2015). After logging in, data users can go to the basic search interface provided by TraitBank by clicking the “Data” tab on EOL’s home page. On the basic search interface, data users select an attribute type, and refine the search by selecting a specific taxonomic group and/or a value for the attribute (Parr et al., 2015) (Figure 4.5). After clicking the “search” button, results will be presented below this button in a dynamic user interface that provides the metadata and data resource for each record.

If the data user decides to download the data, they can click the “download” button on the top right of the search page. The downloaded file will be loaded and listed in the data user’s EOL profile page in the form of CSV (comma separated values) table (Parr et al., 2015). The EOL team chose the CSV format because it is easy to open in common spreadsheet applications such as Microsoft Excel or parse in any programming language

(Parr et al., 2015).

TraitBank Search Results download data

47 results for *body mass* within [Trochilidae](#)

Select an attribute  search

**Refine your search (optional)**

Search within a taxon group

**TAXON GROUP**

For example, enter Trochilidae or Whales. ONLY selections from the drop-down will be used, do not edit the text without selecting a result.

**SPECIFIC VALUE**

For example, heterotrophic.

**Not sure where to start?**

Try one of the following data searches:

- [What are the various shapes of diatoms?](#)

search [Start a new search](#)

Displaying 1 – 47 of 47 records on EOL

	<b>Campylopterus hemileucurus</b> <a href="#">Violet Sabrewing</a>	body mass (max)	12 g	<a href="#">Birds Animal Di...</a>
	<b>Campylopterus hemileucurus</b> <a href="#">Violet Sabrewing</a>	body mass (average)	10.5 g	<a href="#">Birds Animal Di...</a>
	<b>Campylopterus hemileucurus</b> <a href="#">Violet Sabrewing</a>	body mass (min)	9 g	<a href="#">Birds Animal Di...</a>
	<b>Lampornis clemenciae</b> <a href="#">Blue-throated Hummingbird</a>	body mass (average)	7.6 g	<a href="#">Birds Animal Di...</a>
	<b>Phaethornis superciliosus</b> <a href="#">Long-tailed Hermit</a>	body mass (max)	6.6 g	<a href="#">Birds Animal Di...</a>
	<b>Amazilia tzacatl</b> <a href="#">Dusky-tailed Hummingbird</a>	body mass (max)	5.5 g	<a href="#">Birds Animal Di...</a>

Figure 4.5 The TraitBank search result screenshot (Search TraitBank, n.d.)

Besides TraitBank, EOL also provides an Application program interface (API) to any data user who needs to access the data shared on EOL through their own applications (EOL API, n.d.). Using this API to access the data shared on EOL commonly means it is

being used to develop new information systems (e.g., websites, tools, smartphone applications) and the data being embedded into these systems. The API can also be used to access large amounts of data in different formats, so that data users are not restricted by the type and size limitations on TraitBank.

Having introduced the data providers and data mediators who make data sharing occur on EOL, the next section will focus on revealing the data sharing processes by reporting the results of analyzing the EOL JIRA system contents.

#### 4.4 Answering the third question: what are the data sharing processes?

Results from the JIRA system content analyses show that there are two key milestones of ensuring that data sharing from a content partner to EOL happens. The first is that an effective technological connector must be built between EOL and the content partner to enable data to be transferred to and displayed on species pages. Successfully establishing this connector represents the formation of a formal data sharing partnership between EOL and a content partner. In this way, the processes of data sharing could be considered the process of forming the partnership, and includes the contents of the sociotechnical interactions between mediators from both EOL and content partners.

The second milestone is that data sharing processes do not end once the connector is successfully built: building the connector is not a one-time process. Instead, data sharing between content partners and EOL is a series of sustainable processes. It requires human actors to continue making the effort to maintain partnerships between the content partners

and EOL from both technological and human perspectives. Without the appropriate maintenance, the data shared on EOL website would be outdated and might even disappear.

Therefore, answering the research question, “what are the sharing processes?”, refers to revealing how human mediators from EOL and content partners collaboratively built and maintained their data sharing partnerships.

Data sharing processes, that is, building and maintaining data sharing partnerships between a content partner and EOL, are the same at a high level for all the content partners. High-level processes are a series of general processes that any content partner would go through if they want to share data with and display it on EOL. However, at a more specific level, the actual processes vary significantly from one content partner to another based on each partner’s specific situation and needs. Thus, sharing processes are influenced by differences among content partners at both organizational and individual levels.

#### 4.4.1 The general processes of sharing data

A trilogy of interaction activities among social actors was identified as part of the general processes of sharing data between content partners and EOL. The three activities include:

- Preparing social relationships and reaching mutual agreement;
- Developing a data sharing connector;

- Updating the data and/or data sharing connector.

#### 4.4.1.1 Preparing social relationships and reaching mutual agreement

The data sharing processes begin when human mediators from the content partner and EOL start a mobilization conversation about publishing the content partner's data on EOL's species pages. The conversation was usually initiated at a director or manager level. When EOL directors or managers, acting as administrative social actors, are first to identify a data source, they tried to contact people with administrative responsibilities for the data, such as managers of the data source. These EOL staff expressed interest in building a collaborative data sharing partnership and invited the data source to consider becoming one of EOL's content partners. In some other cases, when managers of the data sources identified the EOL data sharing opportunity, they first reached out to EOL by themselves. They introduced their data sources to EOL and expressed interest in sharing those sources with EOL.

These conversations happened online, offline, or both. In some cases, human actors from EOL and the data source already had pre-existing interpersonal relationships; once they started to work on building a collaborative data sharing relationship, their existing relationship was extended to the online environment. However, in most cases, human actors from each side did not know one another so that new relationships were established between EOL and the data sources.

When administrative human actors from both sides were positive about building data sharing partnerships, a data source's role would change to a data provider and potential content partner. Human mediators from this data provider would then work collaboratively with those from EOL to build a partnership, the first step of which being an agreement on data sharing. The items on this agreement include:

- What data can be shared. For example, not all data in a data source can be shared with EOL, only a certain types of data (e.g., image, text, map, etc.), a certain taxon group of data (e.g., plants, birds, etc.), or a certain data creator's data, as agreed by the data provider.
- If the data are or contain elements protected by copyright, what appropriate copyright license should be adopted. For example, photographs are one type of data that are often protected by copyright.
- What appropriate attribution information should be used for the data.
- Whether it is necessary to contact individual data creators. When the copyright license is All Rights Reserved, the data provider and EOL must contact the data creators to ask for permission to share their data on EOL. However, even when the data size is large or the data are taken from academic publication and although the copyright license is Creative Commons, the data provider and EOL might agree to contact the individual data creators to make them aware of what was planned for their data. Usually data creators were satisfied if they were recognized and attributed appropriately by data users.

- What technology should be adopted for transferring data from the data provider to EOL.
- How to divide technology intelligence labor between the data provider and EOL when preparing the data resource that would be transferred to EOL. To be more specific, the data provider and EOL together decided whose and how many technicians should take responsibility for preparing the data resource and how much time and money each side agrees to offer.
- Assign social actors from both sides to take responsibility for different tasks. For example, who should be the administrative and technical contacts for each side.

A few data providers needed to archive a formal copy of the agreement (e.g., Memorandum of Understanding) signed by both sides' administrative social actors.

The process of reaching this mutual agreement helped EOL human mediators to understand the data source in more depth. They need to ensure they had enough understanding of and confidence in the data provenance, that is, where the data comes from should be transparent. Because it is impossible to evaluate the accuracy of such large-scale biodiversity data, the trustworthiness of the data source instead becomes one of the most important evaluation criteria for building a partnership with a data provider.

Another important criterion for establishing a relationship is whether the new data source could provide new data or add value to existing data on EOL (e.g., redesign the way to present data or make complex professional biology data more easily understood by non-

professionals, such as children). If the data is merely a duplicate of that provided by existing content partners, EOL could not accept this data source as a data provider or potential content partner.

After they get the green light from a data provider to build a partnership, either EOL or the data provider's human mediators would register a new account in the content partner registry (i.e., content partner management system). The content partner management system is embedded in the EOL website, and is an important part of the technological infrastructure. This account is used to upload, harvest, publish, and update data.

At the same time, EOL human mediators started to document data sharing processes by creating electronic tickets for each data provider in the JIRA system, which is designed for managing software development and tracking project issues. Each data provider could have more than one of several different types of ticket. Usually, every content partner will first have a "collaboration ticket." As time passes by, in the following processes, one or more sub task tickets might be created, such as "data import," "user feedback," and so on. These tickets are used to record and track the progress, issues, solutions, and any other information related to the data providers' data sharing. These records take the form of internal comments left by EOL human mediators.

The contents of the comments included internal communication among EOL human mediators and external communication between EOL and data provider human mediators. In the comments in JIRA tickets, EOL human mediators copied and pasted or

summarized emails and meeting contents between themselves and data provider human mediators. Only EOL human mediators can view the JIRA tickets since they are only used for internal communication among EOL staff. For example, in the JIRA tickets, EOL SPG directors assigned coordinators and technicians to each content partner by leaving comments; each EOL human mediator reported their progress on their tasks in corresponding tickets, so that everyone would be on the same page.

Usually, the conversation about the agreed items listed above was made at the beginning of the data sharing processes. However, the agreement could be revised at any time, and conversation about the agreement could continue as long as the partnership between a content partner and EOL continues.

#### 4.4.1.2 [Developing a data sharing connector](#)

The essential function of the connector was to effectively transfer the content partners' data to EOL's species pages. The data sharing connector includes two major working components: data resources and EOL content partner registry (i.e., content partner management system). The content partner registry is a data resource management system, embedded in the EOL website as an important part of EOL technology infrastructure. This system was designed and developed by EOL technicians. The major part of the collaborative effort made by both data provider and EOL human mediators when building the data sharing partnership was focused on preparing the data resource.

Data resources are data export files. EOL supports three methods for generating them.

- Extended Darwin Core Archives (DwC-A). This is a relatively new and the EOL preferred method. EOL began to support it since summer 2012. The Darwin Core provides standards for sharing biodiversity information and is one of the most widely deployed formats for biodiversity occurrence data. According to the Taxonomic Databases Working Group (TDWG) (About Us, 2011),

*“The Darwin Core is body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.”*

(Darwin Core, 2015)

By following the guidelines of DwC-A, biodiversity data can be formatted and shared in fielded text formats, such as a set of CSV files.

Although EOL follows the guidelines of DwC-A, it has its own additional requirements. For example, one important requirement is about extensions. Extensions allow new terms to be added to share additional information, so that the data file can be used to serve new purposes (Darwin Core, 2015). Although the DwC-A guide declares that “the extension itself does not have to have a

unique ID” (Darwin Core Text Guide, 2015) EOL requires this so that EOL can keep track of updates to the information belonging to the same data object. (EOL Content Partners: Contribute Using Archives, n.d.)

- EOL XML Transfer Schema. Before EOL supported DwC-A, this method was the most common. According to the EOL website:

*“The basic structure of the EOL XML Transfer Schema consists of a series of taxon elements containing attributes of the taxon as well as one or more dataObject elements providing information about your text descriptions, media files, references, etc. For each taxon, there should be only one taxon element ... Each data object (text descriptions, media files) must be associated with a particular taxon.”* (EOL Content Partners: Contribute Using EOL XML Transfer Schema, n.d.)

Unlike DwC-A, which can include several data files, all data objects need to be in a single file when using the EOL XML Transfer Schema.

- Excel spreadsheet. Data can be submitted in the Excel file format. According to the EOL website:

*“The template consists of a document with 5 sheets representing Media, Taxa, Vernacular Names, References and Agents. Each sheet has two fixed rows which*

*contain the label of the field represented by the column and a brief description of what the field is expected to contain.*” (EOL Content Partners: Contribute Using Spreadsheets, n.d.)

Compared with the other methods, an Excel spreadsheet requires the lowest technical knowledge for generating a data file.

The aim of generating data export files is to ensure data from the content partners can be correctly mapped to its corresponding EOL species pages. There are two layers of mapping. The first is name mapping and pairs a content partner’s data to specific EOL organism pages. By ensuring that each datum has a “Latin binomial,” EOL can associate it with one or more organism names linked to corresponding EOL species pages.

The second layer is chapter mapping. This layer pairs the data object (e.g., text descriptions, media files) to the correct chapter on an EOL species page. There are different chapters of information on each EOL species page: for example, text descriptions could belong to species distribution, habitat, or conservation status, whereas media files could belong to images, videos, or sounds.

EOL usually do not provide server space to host the content partners’ data. If the content partner has already shared the data in an online environment before building data sharing partnerships with EOL, EOL would use the links of the data written in the content partner data export file to access the data and display the data on EOL’s webpage. When the

content partner has not shared the data online, EOL and the content partner need to consider how to digitize the data, find a safe and easily accessible online environment to upload the data to. Sometimes, multiple data files need to be prepared when one data file is not appropriate, for example when the data are for different taxon groups or the data are different types.

No matter what method is chosen to share data with EOL, the following steps must be completed for each content partner to finish preparing the data resource and make the data connector functional.

- The first step is to choose a data sharing method and create a data file or multiple data files. Irrespective of what technology was used and how labor was divided between the data provider and EOL, the goal is to get qualified data export files.
- After the data file is ready, it needs to be validated by using an XML file validator. Any problems need to be resolved by developers from either or both sides.
- After the data file is validated successfully, it is uploaded to the content partner's account.
- A time is scheduled to harvest data, which can sometimes take a couple of days, depending on the amount and size of the data and the server conditions. If problems appear when harvesting data, EOL technicians are ones responsible for solving them.

- When the data from a content partner are successfully harvested, the data need to be reviewed by both sides to verify the right data has been shared, they are displayed correctly, and copyright and attribution information are right. Data review is usually an iterative step and it will continue so long as there are still data and/or technological problems. The data might need to be reharvested several times until both sides are satisfied with the data preview.
- Given that EOL is an open source database, anyone can access the data shared on it and use it. Therefore, all data without an appropriate license need to be excluded. Duplicate data needs to be filtered out as well; different content partners could provide overlapping data because data creators can choose to share their data with multiple online platforms.
- Once data and technology problems are solved, preparation for first-time publication is finished and the data resources are ready to be published.
- Then the last step is to decide when to formally publish data. The SPG director is the specific human mediator who is responsible for making the final decision whether and when to publish. Once the EOL human mediator clicks “publish,” the formal partnership between EOL and the content partner is successfully established. It might take some time, up to a few days, until the data are displayed publicly on EOL.

#### 4.4.1.3 Updating data and/or data sharing connector

Data sharing does not end at the point the data has been successfully transferred from the content partners to EOL. Most databases dynamically change as old data can be updated

and new data can be added; both content partners and EOL value the importance of a data aggregation and integration repository keeping pace with the data providers and of keeping data providers' data up to date.

A few content partners' data are static, meaning the data do not change or update after they are uploaded and displayed in an online environment. But most content partners' data are constantly changing, either becoming stronger and more powerful with increasing amounts and quality, or becoming weaker and fading.

There are two types of updating supported by EOL. The first type is making a regular updating plan during the preparation period for first time publication. Automatic data update mechanisms can be set up in the content partner account through which a content partner's data can be reharvested automatically at a frequency (e.g., once per day/week/month/season) of the content partner's preference. Thus, modifications and/or additions to the previously shared data can be updated to EOL in a timely manner.

The second type of updating is manual updating. When a content partner manager does not want to set up an automatic regular updating mechanism, they can choose to manually update the data whenever the data manager from the content partner is available and willing to do.

Updating helps maintain the partnerships between EOL and content partners as these maintenance activities allow EOL and data providers to build a deeper and ongoing collaborative data sharing relationship.

#### 4.4.1.4 Additional steps

EOL technology infrastructure is able to provide data usage feedback to data providers. Data usage feedback not only makes data reuse on EOL more visible, but also helps data providers have a better sense about how their data are reused. Another additional step of collaboration is that EOL shares data with the data providers so that data interchange/exchange mechanisms are set up between them.

To sum up, the three major groups of sociotechnical interaction activities (i.e., preparing social relationships and reaching mutual agreement; developing a data sharing connector; and updating data and/or data sharing connector) reflect the general steps of collaborative data sharing practices between EOL and its content partners.

#### 4.4.2 The data sharing processes vary with each content partner: the influence of human infrastructure

The general data sharing processes introduced above revealed the tasks included in the three major steps for building partnerships between content partners and EOL. However, when tasks are undertaken by different human mediators, the actual processes vary. One of the interviewees from EOL emphasized that “human” is the most important factor that significantly influences data sharing processes (i.e., partnership establishment processes).

Human mediators from both EOL and the content partners make and implement decisions about how to accomplish the tasks in the three major steps. Based on the perspectives of institutional logics (Thornton et al., 2012), making these decisions and carrying them out are both enabled and restricted by the interrelationships between human actors and the organizations/institutions/communities they belong to. These interrelationships are shaped by how the organization/institution/community they belong to was initiated and developed. In the case of EOL, these interrelationships were reflected from three perspectives:

- Human actors' organizational identities that are available from and activated by the human actors' organizations/institutions/communities;
- The relationships between different human actors (i.e., administrative human actors, technicians, and individual data contributors) in the human actors' organization/institution/community; and
- The data sharing contexts that are available from and provided by the human actors' organizations/institutions/communities.

Therefore, in the case of EOL, decision making and implementation are enabled and restricted by the human actors' organizational identities, the relationships between different human actors, and the data sharing contexts. Figure 4.6 summarizes the relationships between data sharing processes, decision making and implementation, and the interrelationships between human actors and their organizations/institutions/communities.

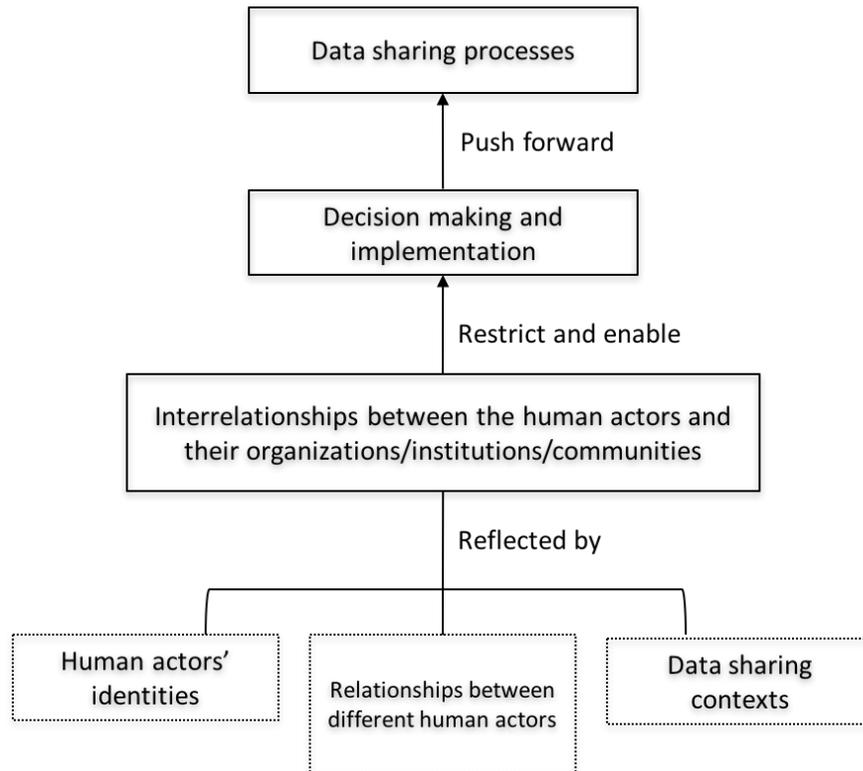


Figure 4.6 The relationships between data sharing processes, decision making and implementation, and the interrelationships between human actors and their organizations/institutions/communities.

Although the general processes of data sharing (i.e., preparing social relationships and reaching mutual agreement; developing a data sharing connector; updating data and/or data sharing connector) were designed by EOL, the data providers (i.e., content partners) usually have a higher degree of control on the processes than EOL. EOL allows the data providers to decide how they want to build and maintain the data sharing partnerships with EOL. The data providers have the power to make final decisions on how they want to share their data. Therefore, the human mediators from the data providers are more likely to play a leading role in causing the variation of data sharing processes.

In the case of EOL, there are two key interrelationships, that is, human actors' relationships with other human actors at individual level, that have significant influence on data sharing processes. The first focuses on the relationships between data managers and data contributors, the second on the relationships between data managers and technicians. The relationships between data contributors and technicians are less evident than the previous two.

#### 4.4.2.1 Relationships between data managers and data contributors

The first interrelationship between individual-level human actors and collective-level data providers was reflected by the relationships between the individual-level administrative human actors of the collective-level data provider (e.g., directors, data managers) and individual-level data contributors (e.g., data creators, data authors). This key interrelationship has significant influence on the entire data sharing process. Three major types of relationships between data managers and data contributors were identified among the content partners:

1. Data manager(s) of a collective-level data provider is/are the individual data contributor(s) who created and contributed the data to the collective-level data provider in the first place. They had full ownership, took full responsibility for the data, and were willing to share their data with EOL. They worked directly with EOL human mediators on building and maintaining data sharing partnerships and may or may not have had a technician work with them.

This type of data manager had already decided where to share their data online before building a data sharing partnership with EOL. This initial online environment became their data's home repository and could be on one of many different platforms, such as a self-service information system built from scratch for only sharing their own data, a pre-established third-party information management platform, an online community, or an information aggregation and integration system built for sharing multiple contributors' data.

In terms of the data sharing contexts, most professional repository content partners and subsidiary content partners fall into this type of interrelationship.

2. Data manager(s) of a collective-level data provider do not or cannot represent the individual-level data contributor(s) who created and contributed the data to the collective-level data provider. They do not own the data, but help manage the data for the contributors by using the information systems and tools provided by the collective-level data provider, which are part of the collective-level data providers' technology infrastructure. They might or might not directly work for (i.e., represent) the collective-level data providers, but could still directly participate in decision making and implementation regarding building the partnership between the collective-level data providers and EOL. Data contributors were not involved in these decision making and implementation processes.

If the data creators chose an all rights reserve license to protect their data, data managers must get permission from them before sharing their data with EOL. However, if a data creator's data is in the public domain or has certain creative commons licenses, s/he might not be aware that collective-level data providers have shared their data with EOL because the data providers do not need to get the permission from the data creators under these circumstances.

In terms of data sharing contexts, some venerable organization, all citizen science initiative, all social media platform, and some education community content partners fall into this type of interrelationship. These collective-level data providers provide online environments, such as online community and data aggregation systems, for data contributors to share data and allow their data to be shared elsewhere.

3. The third type of relationship is that the data manager(s) of a collective-level data provider is/are not the individual-level data contributor(s). The data managers represent the collective-level data providers and do not own the data, but host and manage data for the data contributors. However, the decision to share data could be made not only by the data managers, but also by the key data contributors: managers and contributors collaboratively contribute to building partnerships between collective-level data providers and EOL. Key data contributors were likely the catalysts that increased the possibility and facilitated the process of building successful data sharing partnership between the data provider and EOL.

This type of relationship looks is a hybrid of the previous two types of interrelationships.

In terms of data sharing contexts, some professional repository content partners fall into this type of interrelationship.

#### 4.4.2.2 Relationships between data managers and technicians

The second interrelationship between individual-level human actors and collective-level data providers was reflected by the relationships between the administrative human actors of collective-level data provider (e.g., directors, data managers) and technicians. This second interrelationship has a specific influence on the processes of generating and maintaining the data sharing connector. As discussed above, there are different methods of preparing data export files during the process of building the data sharing connector. How these data export files should be prepared for a data provider depends on the answers to the following questions: 1) how much information technology intelligence the data provider has; 2) what types of technology the data provider agrees to use to create the data export file(s); and, 3) how much effort (e.g., human and technology resources, and time) the data provider is willing to offer to create the data export file(s). The answers to these questions are decided by the relationships between the administrative human actors and the technicians.

Two types of relationships between the data managers and technicians were identified:

1. Data manager(s) of a collective-level data provider is/are the technician(s). They have abilities and skills to deal with the technological demands for storing and sharing data.

Under this condition, the technicians have the same power as administrative human actors. They can make decisions about how they want to collaborate with EOL to build the data sharing connector.

2. Data manager(s) of a collective-level data provider is/are not the technicians.

Under this condition, administrative human actors made the high-level decision to work collaboratively with EOL to build the data sharing connector, and subsequently introduced their technicians to EOL. The technicians only became involved in generating and maintaining the data export files and related technology issues, rather than directly participating in high-level decision making processes.

No matter what relationship the data manager and technicians have with a collective-level data provider, the responsibilities of technicians are very similar: building and maintaining the information systems and tools for storing and sharing data. However, not all data providers have a data manager who is also a technician, or are even able to recruit a technician.

Technicians from different data providers have different levels of expertise and are familiar with different technologies for sharing data. They also have different levels of availability and willingness to work on preparing data export files, much of which depended on their commitment to the data provider and the data providers' commitment to EOL.

In generating the data sharing connector, five ways of collaboration between a data provider and EOL were identified:

- Technicians from the data provider generated the ready-to-use data export files by using either DwC-A guidance or EOL XML transfer schema. EOL technicians provided help whenever the data provider's technicians needed it.
- Technicians from the data provider could only provide the data export files using their own format. EOL technicians then transferred the files by following DwC-A guidance or using EOL XML transfer schema. EOL technicians consulted the data provider's staff about their data files in order to make sure they understood the data provider's data correctly.
- The data provider provided a web service (e.g., API) to EOL and taught EOL what data the web service was able to provide and how to use it. EOL technicians created the script to obtain the data through the web service and from that generated the data export files.

- The data provider could not provide any form of data export files and did not have an established web service. EOL technicians created the script to scrape the data provider's website and obtain the data, and then generated the data export files.
- If the data provider happened to use Lifedesk or Scratchpads platforms to manage their data, established data export tools were already available to automatically create data export files in the EOL preferred format.

To maintain the data sharing connector, technicians collaboratively work with EOL to adjust the existing data sharing connectors based on the changes from the data provider's technology infrastructure and the content shared on that infrastructure. Although EOL could also have these kinds of changes, EOL made their best effort to not influence existing content partners' sharing.

In many other cases, changes happening to a data provider could trigger a single or a series of updates of the data sharing connector. Data managers or technicians from data providers could make requests to update the data shared with EOL. Alternatively, if members of EOL working groups or other human actors in the EOL community noticed any changes made by the data providers, then EOL themselves would adjust the data sharing strategies for data providers. It might not be necessary to engage technicians from data providers to make changes to the current data sharing connector if the changes are to the content (i.e., data), for example after the data sharing partnership has been successfully built, new data are being imported to the data provider's own website or database, or data quality has been significantly improved. Instead, EOL human

mediators would do a force-harvest to get any data that are different from what regular automatic harvesting obtained in the past.

However, if the changes are to the platform structure, server, or data sharing tools, then technicians from EOL and the data provider should work together again to change the type of technology used to generate the data file or improve an existing or develop a new data connector.

Occasionally, the updating of the data sharing connector could be triggered by some special individuals, such as data creators, data users, or EOL data curators who do not have direct working relationships with the data provider. When they spot an issue concerning the content partner's data, or they find their need to share and/or use data are not satisfied, they would contact EOL directly. These individuals' reports prompt EOL human mediators to consider whether they need to take any action to update the data sharing connector or reharvest the data to solve the problem.

#### 4.4.2.3 Relationships between data managers, data contributors, and technicians

Based on the three types of relationships between data managers and data contributors and two types of relationships between data managers and data technicians described above, there could be six possible types of combinations of relationships among data managers, data contributors, and technicians (Table 4.1). The figure in each cell of Table 4.1 visualizes the relationship among these three types of human mediators. The shaded areas in the figures represent which types of human mediators participated in building the

data sharing partnerships with EOL. The existence of these six types of relationships was confirmed by analyzing EOL JIRA system content. Each type leads to a different data sharing story behind building the data sharing partnerships with EOL. In Appendix C, four data sharing stories that represent four different types (cell 1, cell 2, cell 5, and cell 6) of the combination relationships respectively are described in detail to illustrate how the processes of data sharing vary from case to case. These stories share the important similarities with the rest of 32 content partners and refer to how the human actors' organizational identities, relationships with other human actors, and data sharing contexts influence the data sharing processes.

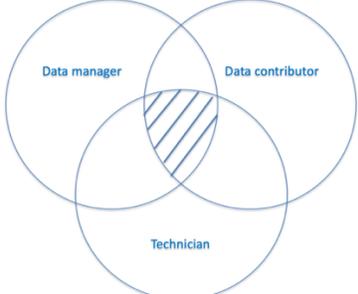
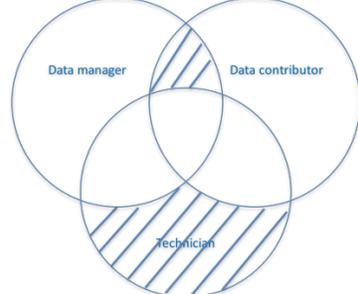
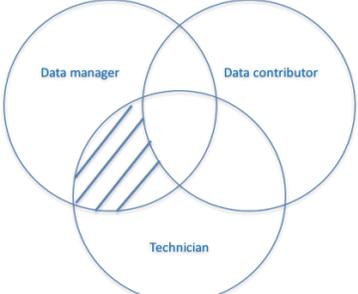
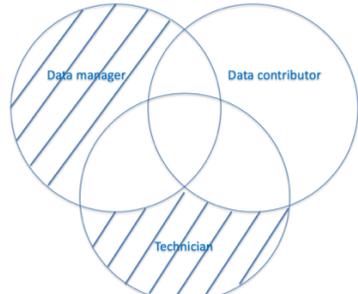
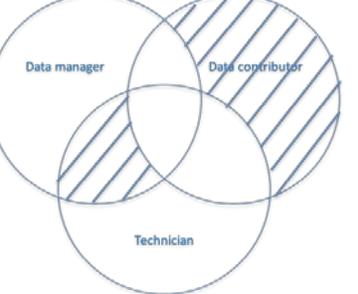
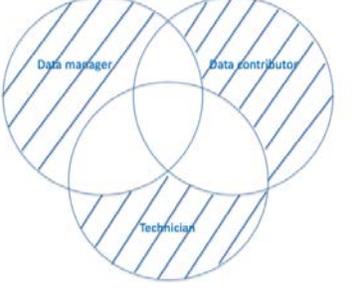
Data managers & technicians Data managers & data contributors	Same person	Different persons
Same person	<p style="text-align: center;">1</p> 	<p style="text-align: center;">4</p> 
Different persons, but data contributors did not work with EOL human mediators on building partnership	<p style="text-align: center;">2</p> 	<p style="text-align: center;">5</p> 
Different persons, both work on building partnership with EOL	<p style="text-align: center;">3</p> 	<p style="text-align: center;">6</p> 

Table 4.1 Six types of combinations of relationships among data managers, data contributors, and technicians.

#### 4.4.3 The number of human mediators and the time for building partnerships

As discussed above, certain types of relationships between data managers and data contributors are more likely to occur within certain types of collective-level data providers with respect to data sharing contexts. These data sharing contexts are available from and provided by the human mediators' organizations/institutions/communities. The

data sharing contexts influence the variation of the data sharing processes through the relationships between different human actors that are shaped by the data sharing contexts. The analyses of the JIRA system content not only revealed what tasks are involved in the data sharing processes, but also revealed how many human mediators were involved in these processes and how much time these human mediators from different types of data providers took to build the partnerships. The findings show that there are noticeable differences between venerable organizations and other types of data providers.

Among the 36 content partners, there were 31 content partner JIRA system contents that contained the specific information of who the human mediators were and how many of them were involved in the collaboration with EOL. Among the 31 content partners, three were venerable organization data providers, 24 were professional repository data providers, two were citizen science initiative providers, and two were social media platform data providers.

For EOL content partners, the results show that venerable organization data providers usually have more human mediators (6–9) than other types of data provider, except one large-scale professional repository data provider who had 8 human mediators. Table 4.2 shows the number of human mediators for the 31 content partners. Among the 24 professional repository data providers, 12 of them had only one human mediator, five had only two human mediators, two had three human mediators, three had 4 human mediators, and the rest had 6 and 8 human mediators respectively. The two citizen

science initiative data providers had two and one human mediators respectively. The two social media platform data providers had three and two human mediators respectively.

<b>Type of content partners</b>	<b>Number of human mediators</b>	<b>Number of content partners</b>
Venerable organization	6	1
	7	1
	9	1
Professional repositories	1	12
	2	5
	3	2
	4	3
	6	1
	8	1
Citizen science initiatives	2	1
	1	1
Social media platforms	3	1
	2	1
<b>Total</b>	<b>53</b>	<b>31</b>

Table 4.2 The number of human mediators for the different types of content partners.

With respect to EOL human mediators, the results also show that the venerable organization content partners worked with more EOL human mediators than most other content partners. However, although the professional repository content partners had a smaller number of human mediators involved in building the partnership, it does not mean that EOL would assign fewer human mediators to work with them compared with any other type of data provider. The correlation between the number of human mediators from the professional repository data providers and the number of EOL human mediators who worked with them is not significant ( $p = .24$ ). A similar phenomenon is also observed for citizen science data providers and social media providers. How many content partner human mediators involved in building the partnership with EOL is not

related to how many EOL human mediators are involved, i.e. there is no correlation between the number of human mediators on each side. Table 4.3 shows the numbers of EOL human mediators and content partner mediators.

Type of content partners	Sample content partner ID	Number of human mediators from content partners	Number of human mediators from EOL
Venerable organization	1	6	6
	2	7	6
	3	9	6
Professional repositories	4	1	4
	5	1	7
	6	1	9
	7	1	2
	8	1	7
	9	1	5
	10	1	3
	11	1	4
	12	1	3
	13	1	2
	14	1	2
	15	1	6
	16	2	3
	17	2	3
	18	2	3
	19	2	3
	20	2	4
	21	3	5
	22	3	4
	23	4	7
	24	4	3
	25	4	6
	26	6	3
	27	8	8
Citizen science initiatives	28	1	2
	29	2	7
Social media platforms	30	2	5
	31	3	9

Total	N/A	84	N/A (Overlapped human mediators from EOL)
-------	-----	----	---

Table 4.3 The numbers of human mediators from EOL and different content partners.

Among the 36 content partners, there are 18 content partners' JIRA system content that contains specific information about when they and EOL started to talk about the possibility of building the data sharing partnership, and when the data sharing partnership was formally and successfully built (i.e., the content partners' data are published on the EOL platform for the first time). Therefore, how long it took these 18 content partners and EOL to establish the partnership can be calculated.

Among the 18 content partners, building the partnerships took one content partner fewer than 20 days, two content partners two months, seven content partners more than half a year, six content partner more than one year, one content partner two years, and one content partner three years (Table 4.4).

How long did it take to build the data sharing partnership between a content partner and EOL?	Number of content partners
Fewer than 20 days	1
Two months	2
More than six months, less than a year	7
More than a year	6
Two years	1
Three years	1
Total	18

Table 4.4 The time the 18 content partners took to build the partnerships with EOL.

These descriptive results about the number of human mediators and the time taken to build partnerships provide background information about the relative number of human

mediators from different types of content partners and EOL, and the time required for building the partnerships. For many content partners' human mediators, building a data sharing partnership takes a relatively long time. The four data sharing stories in Appendix C help to explain how exactly these human mediators collaborate with each other and why building these partnerships takes such a long time.

#### 4.5 Conclusion

The findings in this chapter are summarized regarding the three research questions as follows:

- Who are the data providers on EOL?

The data providers are at a collective level. They are formally called EOL content partners. There are seven types of them: 1) venerable organizations; 2) professional repositories; 3) citizen science initiatives; 4) social media platforms; 5) education communities; 6) academic papers; and 7) subsidiaries.

- Who are the data sharing mediators?

Data mediators include human mediators and technology mediators.

The human mediators include human workers from both EOL and the content partner who worked on developing and maintaining the data sharing partnership. On the EOL side, the core human mediators are members from the EOL Species Page Group and EOL Biodiversity Informatics Working Group, and EOL

contractors. On the content partner side, the core human mediators could be data managers, data technicians, and/or data contributors.

The technology mediators are embedded in the EOL technology infrastructure, including three information systems: the Content Partner Management System, the EOL species page, and the TraitBank system.

- What are the data sharing processes?

Data sharing processes are a series of processes of building and maintaining the authoritative data sharing partnerships between the EOL and the content partners. These processes can also be considered the processes of building and maintaining the knowledge infrastructure.

The general data sharing processes that fits all EOL content partners includes three major steps: preparing social relationships and reaching mutual agreement; developing a data sharing connector; and updating the data and/or data sharing connector.

However, the actual data sharing processes vary with each content partner. The most important influential effect comes from human actors on the content partner side: what their organizational identities are, what their relationships with other human actors from the content partner side are, and what the data sharing contexts they come from are.

## 5 Case Two – The CyberSEES Project

### 5.1 Overview of the findings

Chapter 4 focused on revealing the data sharing contexts and processes in the case of EOL's knowledge infrastructure where a large-scale, online, biodiversity data aggregator repository is its major product. Its data sharing contexts are reflected by the different types of collective-level data providers (content partners) and its organizational structure. The collaborative efforts made by human mediators from EOL and the data providers supported by technology mediators reflected the data sharing processes. The data sharing contexts and processes provided a good explanation of the data sharing practices on EOL for the benefit of both research and public communities.

However, the first case only tells half of story about data sharing, that is, sharing from collective-level data providers to EOL. Data sharing processes are not comprehensively understood without the other half of the story, sharing from individual-level data providers to collective-level data providers. Understanding this process reveals data sharing processes that occur before data providers share data with EOL.

Therefore, to tell the other half of the story, this dissertation investigates a second case: an NSF funded distributed cyberinfrastructure project named CyberSEES. A citizen science project, Biocubes, was developed by the CyberSEES project members as the vehicle for creating fresh citizen science data and studying the infrastructure development and design for sharing citizen science data (CyberSEES project description, 2015).

CyberSEES encourages anyone, especially those in the education community, to

participate in Biocubes, where participants are instructed to “(e)xamine one cubic foot of space and discover and report all the living things that are found within it” (Biocubes, 2015). Here “report” means sharing the observation data of living organisms in an open access online environment so that researchers in the research community as well as the public community can access these observation data. CyberSEES project members chose an online environment provided by iNaturalist ([inaturalist.org](http://inaturalist.org)), a citizen science initiative platform and online community of naturalists, upon which Biocubes’ participants are encouraged to register and share their data. After becoming an iNaturalist community member, participants can report all the living things that are found within one cubic foot of space.

As in the first case, in order to investigate how the data in this citizen science project were collected and reported to a collective-level data provider and then shared on an aggregator repository, the three research sub questions are asked:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

The data for this dissertation were collected and analyzed to answer these three questions from the three different types of evidence:

- Artifacts (e.g., website content, such as Biocubes web pages, data records web pages, etc.);
- Documentation data (e.g., meeting notes, EOL JIRA system content);

- Field notes and photos when observers (a senior researcher and the author of this dissertation) participated and observed a Biocubes training workshop organized by CyberSEES project members. The training workshop was held in Florida Atlantic University's Harbor Branch Oceanographic Institute between January 23–25, 2015. The targeted participants of this workshop were science educators from the education community.

## 5.2 Answering the first question: who are the data providers?

Biocubes can be considered a collective-level data provider whose data sharing contexts and mediators were prepared by CyberSEES. Compared with the large size of the biodiversity data aggregated from collective-level data providers in the EOL case, Biocubes can be considered a small-scale biodiversity data source whose data are created and shared by individual-level data contributors. This dissertation focuses on not only collective-level data providers, but also individual-level data providers. These individual-level data providers 1) collected the data; 2) were willing to share their data publicly; and, 3) made the data available in an online environment. If someone only meets the first condition, s/he is solely a data creator. However, if someone meets the first and second conditions, but does not make the data available in an online environment him/herself, this person can still be considered a data provider since others can help with uploading data to a publicly viewable online environment.

Making data available in an online environment requires the data to be transferred from an offline environment to an online environment. At this point, they become visible data

points for potential data users. The online environment where the data originate, that is, the location where the data are first uploaded, is provided by the data's *home repository*. The home repository can exist in the form of a professional repository, a social media platform, or a citizen science initiative, for example.

The home repository of Biocubes' data is iNaturalist, a citizen science initiative website that was built in the form of a social network site of naturalists. CyberSEES researchers created a project page for Biocubes on iNaturalist (Biocubes, 2015). This page is not only one web page but looks like the home page of a website designed for Biocubes, although it is actually embedded in the iNaturalist website (Figure 5.1). The Biocubes site's multiple pages supports different functions, each of which focuses on managing biodiversity observation data. CyberSEES researchers use these pages to manage all the Biocubes data uploaded to iNaturalist by project participants. This "website" also supports a small online community of Biocubes project managers and participants.

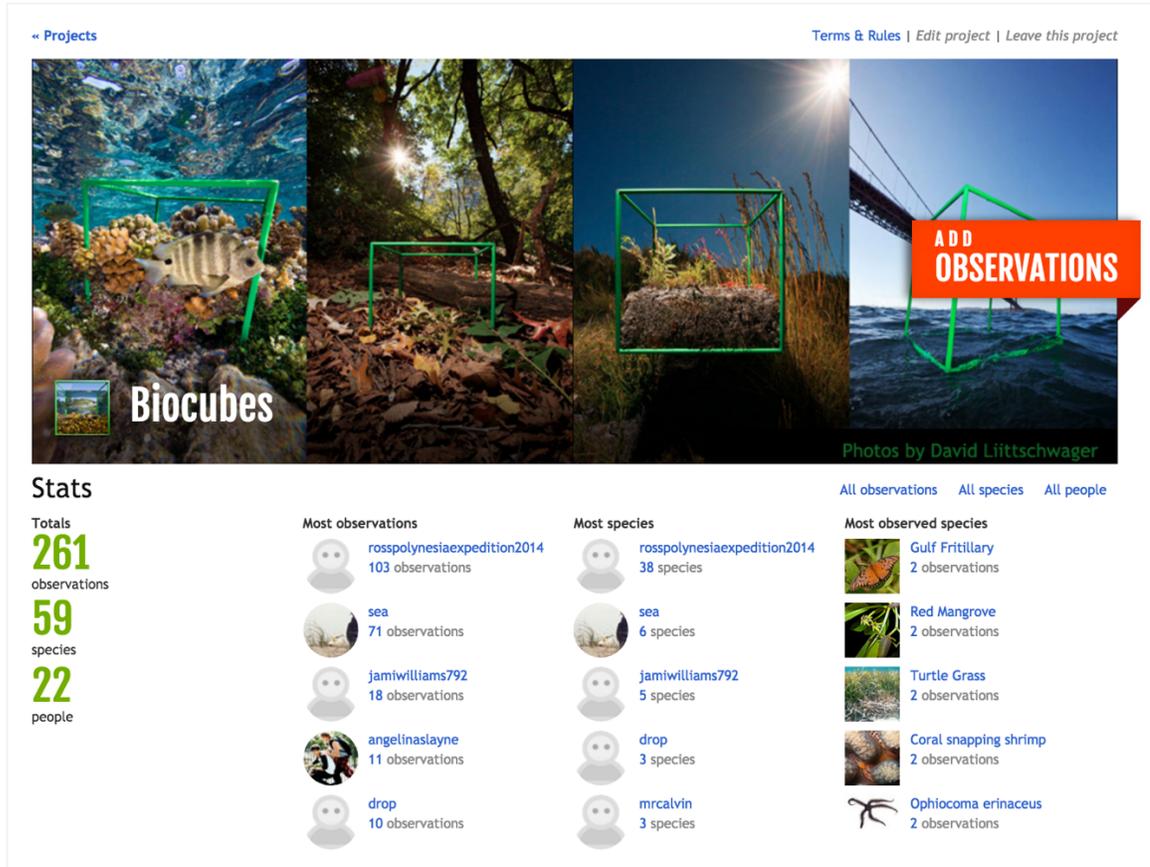


Figure 5.1 Screenshot of the iNaturalist Biocubes project page (Biocubes, 2015).

According to participant observations of the training workshop, the follow-up after the training workshop, and the website contents, there is incongruity between who collected data in the offline environment (i.e., individual-level data creators) and who submitted it to the online environment (i.e., individual-level data providers). In other words, data creators and data providers do not appear to always be the same people: the visible individual-level data providers in iNaturalist only represented some of the Biocubes participants who collected data in the offline environment (i.e., individual-level data creators). Many individual-level data creators were invisible in the online environment.

### 5.2.1 Visible data providers in the online environment

Analyses of the contents on the iNaturalist Biocubes project page revealed that there were 59 biodiversity observation data uploaded by 11 participants (individual-level data providers) of the Florida Biocubes training workshop. Each of these 11 participants registered on iNaturalist and became a community member. Observation data are linked to the participants' iNaturalist user account page. Among these 11 participants' account pages, seven of them provided their real names, three of them provided personal photos (self-portraits), but only one added a short description of her expertise and interest in nature. Since the data providers provided very limited personal information, it is hard to identify their authentic identities.

### 5.2.2 Data creators in the offline environment

CyberSEES project members recruited Biocubes workshop participants with help provided by staff from the Smithsonian Marine Station and The Centers for Ocean Sciences Education Excellence in Florida. Recruitment methods included advertising the workshop via a newsletter, social media posts, and sending email invitations to local science education community members. They also talked to individual educators in person and encouraged them to participate in the workshop. There were a total of 20 people registered to attend the workshop, 13 of whom participated.

Biocubes training workshop participants, representing formal or informal science educators, were the core data creators in the workshop. Table 5.1 shows their organizational identities.

Participant pseudonym	Identities	Gender	Grade levels
Brooklyn	Science teacher	Female	Middle school
Ava	Science teacher	Female	High school
Chloe	Aquarium educator	Female	N/A
Abigail	Education consultant	Female	N/A
Camila	Science teacher	Female	Middle school
Sophia	Science teacher	Female	Middle school
Samuel	Science teacher (prospective)	Male	Middle school
Jacob	Aquarium educator	Male	N/A
Scarlett	Science teacher	Female	High school
Layla	Instructional partner	Female	High school
Samantha	Science teacher	Female	High school
Julia	Aquarium educator	Female	N/A
Jessica	Science teacher	Female	Primary school

Table 5.1 The data creators' organizational identities.

The eight science teachers, who taught in Title I public schools, accounted for the majority of participants. The education consultant and instructional partner used to be middle school science teachers but worked in management roles when they participated in the workshop. Whereas the instructional partner worked for a public school, the education consultant worked with a variety of schools, including both public and private schools. The three educators from aquariums were responsible for developing and delivering educational programs and presentations in their aquarium. The science teachers represent educators who focus on formal science learning, while the aquarium educators represent those who focus on non-formal and informal science learning.

From observing and talking with participants, the observers learned their motivations and reasons for participating in this workshop. Their motivations and reasons can be categorized based on four different types of motivation developed by Deci and Ryan: (1) intrinsic motivation, in which people engage in a certain activity for its own sake; (2)

external regulation, in which people's behavior is regulated by external demands, rewards, or concerns about potential negative consequences; (3) identified regulation, in which people identify with the importance and value of a certain behavior, and (4) amotivation, in which people cannot connect the behavior with any purpose or expectations (Deci & Ryan, 1985; Ryan & Deci, 2000).

For the workshop participants, their motivations to participate in this workshop and the Biocubes project include:

- Intrinsic motivation. All the participants had a personal interest in biology, biodiversity, ecology, and conservation. They love nature and enjoy being outdoors and getting close to it. They showed strong interest in the concept of one cubic foot and the examples of previous Biocubes done by others. They felt that the major workshop activity—collect Biocubes' data with professional scientists and photographers along the shoreline in Fort Pierce—was very appealing and novel. They were excited about doing this kind of activity.
- External regulation. For the middle school and high school science teachers, participating in this workshop can be counted towards their professional development requirement. Professional development provides teachers “opportunities that will help them enhance their knowledge and develop new instructional practices” (Borko, 2004, p. 1). Completing professional development has been required by Florida law since 1998 as an important

condition of renewing school teachers' professional educator certificates<sup>1</sup> (Rubio & Pickens, 2008) and teachers earn inservice credits or points by participating in different professional development activities. A teacher must complete a certain number of inservice credits in their specialization area(s)<sup>2</sup> (Rubio & Pickens, 2008). For the remaining participants, no significant external regulation motivation was found.

- Identified regulation. No matter whether the participants had a significant external regulation motivation or not, there was a consensus that participating in this workshop was important. This importance can be reflected in three ways.

First, although they had some previous experience doing hands-on scientific projects in classroom or non-classroom environments, the participants had little experience of introducing citizen science projects to science learning. They had, however, realized that citizen science was a new and promising teaching practice with great potential to increase students' interest in science and science learning outcomes. The participants also highlighted that citizen science could provide students with unique opportunities to collaborate with real scientists and contribute to real scientific projects.

---

<sup>1</sup> Section 5, ch. 86-156, Laws of Florida (1986) (effective July 1, 1988); former § 231.24(2)(a)1., Florida Statutes (1988).

<sup>2</sup> Section 1012.585(3)(a), Florida Statutes (2007). Section 1012.585(3)(c), Florida Statutes (2007); Bureau of Educator Certification, Florida Department of Education, Florida Educator Certification Renewal Requirements (2005) [hereinafter Educator Certification].

Second, the participants agreed in the importance of providing more outdoor activities like the Biocubes project to students. It could be helpful to improve their interest in and awareness of nature and biodiversity around them and that influences their daily life.

Third, besides gaining inservice credits, the participants identified with the importance and value of participating in this workshop. They expected to broaden their horizons to a potentially new way of teaching and improve their knowledge and skills so that they might be able to implement similar citizen science projects. By the end of the workshop, they successfully created concrete Biocubes implementation plans to fit their daily teaching based on what they learned in this workshop.

- Amotivation: The participants' three types of motivation described above show that the participants had clear interests, goals, and expectations about participating in this workshop. Therefore, they had little amotivation.

Some of the CyberSEES project members including biologists, social scientists, and professional photographers also participated in data collection. They will be introduced in more detail in the following section.

### 5.3 Answering the second question: who are the data sharing mediators?

In the case of CyberSEES, two layers of sharing were specified: 1) sharing data created by non-professionals on the home repository (i.e., iNaturalist); and, 2) sharing data created by non-professionals on the aggregator repository (i.e., EOL). The data sharing mediators', who include both human and technology mediators, responsibility is to make both layers of data sharing occur.

### 5.3.1 Human mediators

Three groups of human mediators were identified as part of the two layers of data sharing. The first ensure data sharing occurs on the home repository. The second and third groups enable data sharing on the aggregator repository.

#### 5.3.1.1 Human mediators ensure the data are shared on the home repository

The key responsibility of the first group of human mediators is to ensure the data collected by the individual-level data creators are shared publicly on the home repository (i.e., iNaturalist). These human mediators were CyberSEES project members, including biologists, education specialists, professional photographers, and citizen science researchers/social scientists from three different research and education institutions: the Smithsonian Institution, National Geographic, and the University of Maryland.

These human mediators are a group of professionals who have expertise in different fields. They share the same belief that enabling the widespread sharing of biodiversity data created by non-professionals in open access environments is important for both science progress and education of the general public. However, neither the human mediators or their institutions alone were capable of fulfilling the goals of promoting this

belief, giving rise to the need to thoroughly study current social and technological conditions that give rise to open sharing in order to develop powerful technology infrastructure to support it. Therefore, they came together and built a formal collaboration relationship to achieve these goals, the first step of which was writing an NSF proposal, which was ultimately accepted and funded (i.e., CyberSEES project).

In order to reach their goals, the human mediators first designed an actual scenario, the Biocubes project, to support non-professionals in collecting biodiversity data in an offline environment. In this project, they developed a data sharing protocol and encouraged participants to share their data in an open access environment. They chose the existing platform iNaturalist because it was developed for supporting public sharing of biodiversity observation data created by anyone.

After finishing the preparation of the scenario and the online environment, the human mediators organized several Biocubes training workshops in different US cities to encourage and recruit non-professional, individual-level data creators/provider to participate and share data. The human mediators played the role of workshop facilitators. A senior researcher together with the author of this dissertation were among the human mediators and observed one of these workshops in Florida.

Most CyberSEES project members participated in the Biocubes training workshops held in Florida in early 2015. Table 5.2 shows their organizational identities. Besides the two

observers, there were a total of 11 workshop organizers/facilitators, including one temporary helper.

Workshop organizer pseudonym	Identities	Gender
Jocelyn	Education specialist from SI	Female
Tessa	Education specialist from SI	Female
Charles	Biologist from SI	Male
Simon	Biologist from SI	Male
Melanie	Biologist from SI	Female
Lydia	Education specialist from SI	Female
Daniel	Photographer from National Geographic	Male
Alex	Photographer assistant from National Geographic	Male
Shane	Project coordinator from SI (Helper)	Male
Alison	Social scientist from University of Maryland	Female
Rebecca	Social scientist from University of Maryland	Female

Table 5.2 The Biocube project organizers' organizational identities

### 5.3.1.2 Human mediators ensure the data are shared on the aggregator repository

Two groups of human mediators ensured that the data were shared through the home repository, iNaturalist, to the aggregator repository, EOL. One group focused on building a reliable data sharing channel, whereas the other group focused on improving the data quality on the home repository so that the data would be accepted by EOL.

#### *Human mediators for building the data sharing channel*

The group of human mediators that focused on facilitating the data sharing from iNaturalist to EOL did so by building the reliable data sharing channel between the two platforms. By transferring the data from iNaturalist to another platform, the data are presented to a wider audience and therefore more potential data users who find data useful to them.

These human mediators included administrative human actors and developers from iNaturalist and other platforms. When iNaturalist built formal data sharing partnerships with other platforms, they become data partners. Formalizing these data partnerships guarantees the reliability of the data sources for potential data users and the trustworthiness of the sharing channel and environments for data creators.

iNaturalist has built formal data sharing partnerships with different collective-level partners, including research institutions, professional data repositories, conservation organizations, citizen science communities and projects, and so on (Loarie, 2015). EOL is one of iNaturalist data partners and, in turn, iNaturalist is an EOL content partner. These mutual data sharing partnerships provided a unique research opportunity to investigate data sharing practices from the origin of the data, to the initial sharing of the data in a first open access online environment, home repository iNaturalist, and to continued data in a second open access online environment, aggregator repository EOL.

Given this, the specific human mediators could be identified by analyzing EOL JIRA system content that was created just for iNaturalist. The human mediators appearing on within JIRA system content include two directors from iNaturalist (one also being an iNaturalist technician), three members from the EOL SPG, and one EOL technician.

*Human mediators for improving data quality*

Building the reliable data sharing channel between the home repository and the aggregator repository is one key precondition of sharing data between the two. However, this channel alone is not enough to ensure data are shared. Another essential condition is that the data must reach the standard required by the aggregator repository; iNaturalist and EOL agreed that EOL would accept and aggregate iNaturalist data only if it reached “research-grade.” The majority of data aggregated by EOL are collected by professional researchers and is automatically considered “research-grade.”

Therefore, another group of human mediators is necessary to help improve the data quality on the home repository, this group being iNaturalist community members.

iNaturalist provides features to support community members help each other review data and improve its quality. For example, members can check whether the identification of an organism is correct, suggest an identification at a more specific taxon level, or whether the observation location, time, and description are plausible. With this help, some data becomes research-grade data and it can be transferred to the aggregator repository via the established data sharing channel. Therefore, only Biocubes data that reach research-grade level on iNaturalist can be aggregated by EOL and displayed on their pages. The process of iNaturalist community members helping data reach research-grade is introduced in the next section.

### 5.3.2 Technology mediators

Technology mediators were used by human mediators to enable data sharing. As previously mentioned, there were two layers of data sharing in this case. The

corresponding two layers of technology mediators were identified in Biocubes. This section focuses on the first layer since the second layer were introduced in the previous chapter. The first layer of technology mediators includes three information systems embedded in the iNaturalist infrastructure:

- Data entry system;
- Organism observation page; and
- Data exporting system.

#### 5.3.2.1 Data entry system

iNaturalist's data entry system connects the data creators/providers to the iNaturalist platform. Anybody who has Internet access can register and create a user account, and directly add biodiversity observation data as an iNaturalist community member via their personal computers (i.e., iNaturalist webpage) or smartphone (i.e., iNaturalist app). The user account can be either an individual account that represents a single person or a joint account that represents a group of people (e.g., a research team). iNaturalist assumes the person who adds data is the one who observed the organism and therefore terms them "observer." However, this person does not have to be the actual observer who goes out in the field and collects the data: those responsible for uploading data to iNaturalist could be data managers, for example.

To add observation data on PC, for example, an iNaturalist member, acting as an observer, clicks the "add" tab after s/he logs in; a new page for entering data will pop up. This page allows the observer to add only one datum or a batch of data (Figure 5.2).

Irrespective of their choice, the data entry system asks three questions: “What did you see?”, “When did you see it?”, and “Where were you?”. The observer’s answers to these questions constitute the most basic elements of biodiversity observation data. The data entry system also encourages and supports adding media (e.g., photos, audio), a text description, and tags about the organism(s). The observer can also customize the data fields by adding any additional new data fields appropriate for that observation.

However, none of these pieces of information is an essential element to create an observation datum or batch of data. An observer can choose to provide as much or as little information as they want. Even an empty observation datum can even be created, as long as s/he clicks the “save observation” button before leaving the data entry page; in this case, the datum will be given “something” as the name of this organism, and “somewhere” as the location of the observation. The iNaturalist member who adds an empty entry can always come back to edit it by adding more meaningful information or just delete it. The purpose of an empty entry encourages observers to record their observations without feeling forced to provide any information they do not have or do not want to share.

**Add an observation** [Add: Batch](#) · [From list](#) · [Import](#) · [From photos](#)

What did you see?  ID Please?  **Where were you?**

Was it captive / cultivated?

When did you see it?

(GMT-05:00) Eastern Time (US & Cai)

e.g. "2016-04-30 18:03:56", yesterday at 4pm

**Description**

**Tags** *Comma-separated, please*

**More fields**

Add a field

**Add media**

**Select one or more photos**

No file chosen

Sync obs. w/ photo metadata?

We also support [Flickr](#), [Picasa](#), and [Facebook](#) for image hosting.

Map

Lat: Lon:

Acc (m): Src:

Click the map to add a location.

Google Map data ©2016 Imagery ©2016 NASA Terms of Use

Change geoprivacy

Figure 5.2 Screenshot of the data entry system on iNaturalist.

### 5.3.2.2 Organism observation page

For every datum that is successfully added by an observer, a corresponding, publicly viewable, organism observation page will be generated automatically on the iNaturalist platform. Figure 5.3 shows an example of an organism observation page. Besides observation information submitted by an observer, each organism observation page provides other information that might be helpful for data users to decide whether this observation datum suits their needs, for example:

- The link to the observer's profile page, the device used by the observer to add the data, and the copyright adopted by the observer for both the media element and the entire observation data. In other words, data users can see who added the data and using what device(s), which might be helpful to evaluate the trustfulness and the quality of the data;
- The data quality assessment results calculated by the iNaturalist computer algorithm, which is discussed below. Data users can judge the value of the data based on the data quality assessment results;
- iNaturalist community members' interaction with the data, the data observer, and with other community members, such as suggesting a name for the organism or leaving comments to the observer or other members. Data users might find this conversation is helpful to get to know the organism better.

In addition, if people visit this organism observation page and find the organism fits their interests or use, the page supports the user marking this observation as some of their favorite data (provided they are a logged in iNaturalist member) or add this observation to a project (provided the project already has a page on iNaturalist). If they would like to learn more about this observation, they could leave comments on this page to ask the observer if s/he could provide more details.

« Your observations Previous Next

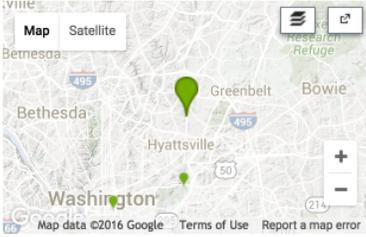
**star magnolia** (*Magnolia stellata*) · Observed by yrhe · March 16, 2016 · 05:34 PM EDT

[Edit](#)
[Copy](#)
[Delete](#)
[Add to favorites](#)
[Identify](#)
[Add to project](#)
[Share](#)



Photo © yrhe, some rights reserved  
[Add more photos](#)

Added: Mar. 17, 2016 17:29:18 -0400  
App: iNaturalist Android App  
[Add/edit more fields](#)



Location: 7346 Baltimore Ave College ... (Google, OSM)  
Places: I-495, Prince George's, US-MD, US Eastern States, US, North America, NA [More...](#)  
Lat 38.9805947, Lon -76.9388742  
Accuracy: 20m  
Geoprivacy: open [Edit](#)  
[Hide details](#)

Your ID:  
**star magnolia** (*Magnolia stellata*)  
[Remove](#)

**Identotron**

**Data Quality Assessment**

Community-supported ID?	No	0 people agree	0 people disagree
Community ID at species or lower?	No		
Community can confirm/improve ID?	Yes	What do you think?	No
Date?	Yes		
Georeferenced?	Yes		
Photos or sounds?	Yes		
Is the organism wild/naturalized?	Unknown	What do you think?	Yes / No
Does the location seem accurate?	Unknown	What do you think?	Yes / No
Does the date seem accurate?	Unknown	What do you think?	Yes / No
Appropriate?	Yes	Inappropriate? <a href="#">Flag this observation</a>	
Quality grade	<b>Needs ID</b>		

[Hide details](#)

**Comments & Identifications**

 Your ID: **Flowering Plants** (Phylum: Magnoliophyta)

Posted by you about 1 month ago

I'm not certain but this could be a Star Magnolia (*Magnolia stellata*) or a related hybrid.

Posted by [redacted] about 1 month ago

 @ [redacted], thank you very much! I will updated the identification name of this plant based on your suggestion, and to see how other people think. :)

Posted by you about 1 month ago

 Your ID: **star magnolia** (*Magnolia stellata*) [Remove](#)

Posted by you about 1 month ago

[Add a comment](#) [Add an identification](#)

[Save comment](#)
[Preview](#)
 [Subscribe to this observation](#)
 [Mark as reviewed](#)

Figure 5.3 Screenshot of an organism observation page on iNaturalist (He, 2016).

### 5.3.2.3 Data exporting system

iNaturalist not only shares data by exhibiting observation records on web pages but also allows any data users to directly download it. Except media elements, such as

photographs and audio, and comments left by other iNaturalist members, most data elements are downloadable through a data export system embedded in the iNaturalist infrastructure. It allows data users to download either one specific observation datum or a batch of observation data.

Since iNaturalist makes each observation datum available through the data search feature, downloads can be initiated from this feature. The data users need register on iNaturalist and log in before they download data. After log in, they first need to go to the basic search interface by clicking the “Observation” tab on website header.

In addition to the basic search interface, the data users can type in a specific taxonomic group name or a species name and/or the location, then click the “Go” button. Similarly, they can click the “Filters” button, then an option panel will pop out, as shown in Figure 5.4. On this panel, data users can customize their search by selecting: 1) specific characteristics of the data (e.g., data elements, description/tags) they are interested in; 2) pre-decided taxonomic categories rank values and a way of sorting based on the data users’ preference; and/or 3) the specific date or a range of time of the observation. Search results automatically appear.

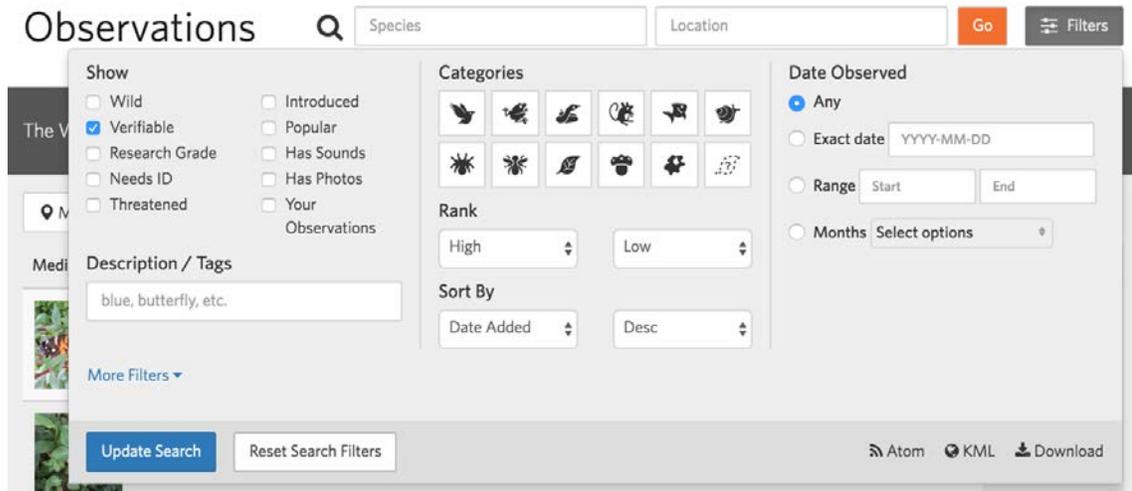


Figure 5.4 Screenshot of the basic search and the Filters panel on iNaturalist (Observations, n.d.).

When data users decide to download the data, they click the Filters button again, and click the “Download” button on the right bottom of the Filters panel. Data users are directed to a web page for exporting data (Figure 5.5). Going from the top to bottom of this page, data users will first see the query they created on the Filters panel, unless they choose to first update their query by using the Filters panel presented here. This Filters panel requires the data users to specify a taxon, place, users, and project to create a new query, or directly specify a search query that has already been created before.

Below the Filters panel, data users will preview the data matching their filtering request. The preview includes basic metadata of these data, including media elements, organism identity, observer identity, data observed, place observed, and how many identification(s) these data have logged from iNaturalist community members.

Below the data preview, data users can customize what columns they want to have for each of the datum in the database: each column represents a characteristic or element of an observation. The columns are categorized into five general groups.

The first group is called “Basic,” which includes: a unique organism ID name decided by the data observer, the iNaturalist community member, and iNaturalist itself; the specific time and date of the observation; the data observer’s account name; the data quality grade, license information; the URL for the datum page, media elements, tag, description, number of agreements/disagreements from iNaturalist members; and, what device(s) were used to upload each observation datum.

The second group is called “Geo” and includes latitude, longitude, positional accuracy information, privacy setting of geographic information, the name of places if possible (e.g., town, county, state, and country), and so on.

The third and fourth groups are both for taxon information, and are called “Taxon” and “Taxon Extra” respectively. “Taxon” includes the name of the species that is given by the data observer, as well as the scientific name and common name given by the iNaturalist. “Taxon Extras” include the scientific name at every taxon level.

The last group, “Observation Fields,” includes the fields each data user set up in their iNaturalist user account when they uploaded their own observations, such as a citizen science project name. After the data user has finished choosing the columns, they click

the “Create export” to download a CSV file that also appears at the top of the web page. Data users can also choose to receive data set via email.

In addition to directly downloading the data set, iNaturalist also offers an API to allow data users from research and public communities to retrieve different forms of data that are uploaded to iNaturalist (Ueda, 2016).



## 5.4 Answering the third question: what are the data sharing processes?

Sharing processes focus on how to create shareable data on the home repository and then transfer the data from the home repository to the aggregator repository.

### 5.4.1 Share to home repository

The processes of sharing data to a home repository starts when a data creator collects and prepares data and ends when the data are successfully uploaded and displayed publicly on the home repository. Participating and observing the Biocubes training workshop revealed the details of each step of these data sharing processes.

The workshop ran from the evening of January 23, 2015 to the afternoon of January 25, 2015. During the first evening of the workshop, the project organizers introduced basic information about the workshop and Biocubes. They introduced themselves to the participants, and let participants introduce themselves to each other. Then they all had dinner together so everyone had the opportunity to get to know each other.

On the second day, the workshop organizers taught the participants how to do a real biocube step by step. Then the workshop attendees, including participants and facilitators, were divided in three groups, each trying to do a biocube. The participants did the primary work of collecting and sharing biocube data but received help from the workshop facilitators. The third day of the workshop focused on asking questions, summarizing biocube experience, and discussing biocube implementing strategies in real-life science learning environment.

At the end of the third day, the workshop participants were provided with an opportunity to do another two biocubes for anyone who would like to practice the complete process of collecting and sharing Biocubes data again.

The remainder of this section will report the content of each step in the processes, based on the three cubes done on the second day of the workshop.

#### 5.4.1.1 The origin of data in the offline environment

According to observations, three major steps in the process of collecting Biocubes data have been identified:

##### *Preparation (Indoor)*

The workshop organizers first illustrated how to do a biocube in different natural environments by playing a few short videos of previous Biocubes field trips. One organizer then provided an introduction of what iNaturalist is as well as detailed guidance about how to submit their Biocubes data to iNaturalist.

After playing the videos, the organizers offered a biocube kit to each participant, and taught them how to assemble it by themselves. Each biocube kit includes 12 one-foot lengths of 1/4-inch aluminum tubing that are painted green and 24 pieces of copper wire, about 4 inches long (Figure 5.6). Once the participants built their hollow green biocube, under the organizers' guidance the participants discussed the meaning of the space in this

one cubic foot and thought about what an interesting biocube site in an outdoor environment might be.



Figure 5.6 The Biocubes project participants were using biocube kit to build biocubes.

#### *Putting biocube in site (outdoor)*

The 13 participants were divided into three groups; each took responsibility for collecting one biocube data set. Therefore, there were a total of three biocubes, each assigned a unique ID: cube\_1, cube\_2, and cube\_3. Each group was led by one workshop organizer and was first asked to choose a site along a seacoast. The organizer helped to provide necessary guidance and answer questions, but did not make any decisions for the participants. The first group chose a site on the wrack line, the second group chose a site among mangrove roots in shallow offshore water, and the third group chose a site in a swamp.

*Collecting and sorting offline Biocubes data (outdoor)*

After each group put their cube in the selected site, they filled out a hard-copy “observation sheet” with a pen or pencil, recording the general conditions of the biodiversity and natural environment within and around the cube. They not only paid attention to the stable lives inside and outside of the cubes, but also to any moving lives that came in and out.

They took photos of the undisturbed cube from different sides and angles. After observing the biocube and the environment within and outside of the cube for about 10 minutes, the participants started to collect the cubes for extraction. By using insect nets, sucking-type aspirators, dip nets, plastic vials and jars, and so on, they caught animals that were in or passing through the cubes to help maximize the number of specimens taken. They then used digging tools, such as shovels or trowels, to extract anything in the cube, such as soil, water, and any other living creatures, putting them in containers, such as buckets and plastic bags.

The participants brought the cube content into a lab. They transferred the cube content carefully to big white trays, and used plastic spoons, soft tweezers, and pipettes to find and catch organisms within the biocube content. The participants tried to group similar organisms together identify them using both offline and online materials (e.g., marine biology books, Google) with the organizers’ help. Microscopes and books of marine life and organisms living in coastal zones were also offered to help.

#### 5.4.1.2 The origin of data in the online environment

After collecting Biocubes data in offline environment, one major step in the process of sharing Biocubes data in the online environment has been identified:

##### *Digitizing and sharing Biocubes data (indoor)*

After the organisms had been sorted into groups, the participants took photographs of them using their own smartphones. In order to help the participants be able to take photos of small organisms and take good quality photos, the workshop organizers provided them with a smartphone lens kit, mini tripod, petri dish, and small pieces of velvet or white cloth/paper so that the participants could put the organisms against a contrasting background to get the best images.

Taking photos of the static organisms was much easier than taking them of the moving organisms, for example an ant. When a participant wanted to take a photo of a small ant, she first carefully moved the ant into a petri dish. But because the ant kept moving fast in the dish, despite the participant's valiant efforts, the photos she took were blurry. She asked advice from one of the organizers, a professional photographer, and got the suggestion that she could put the ant into the fridge for a little while. She did so, and when she took the ant out, the ant was frozen and could not move, so she could take a few clear photos of it (Figure 5.7). After a few seconds, the ant came around and started to moving around again. This is a prime example of how the participants digitized the offline Biocubes data.



Figure 5.7 A Biocubes project participant was taking a photo of a frozen ant.

After the participants took photos of any organisms they were interested in from the cube, they selected a few of them and uploaded them to iNaturalist with additional information, such as the date and location of the observation and the participants' best guess at the organism's identification which could be at any taxon level.

Before a participant could upload the digitized Biocubes data to the iNaturalist, s/he needed to download the iNaturalist mobile app to their smartphone, register, and create a user account. If a user account could not be completed on the smartphone, a participant could visit the iNaturalist website on a PC and register there. Once the user account was created, the participant logged in to the smartphone iNaturalist app.

Once the participant was ready to add Biocubes data to iNaturalist (e.g., the ant datum), the participant opened the app on her smartphone, logged in, then clicked the “add” tab on the home page. A new page for entering the data would pop up as described above. On this page, the workshop organizers required the participant to upload at least one photo, type in the identification name, the observation time and geolocation, chose the project name (i.e., Biocubes), and finally type in the unique cube ID. The project name and unique cube ID were new data fields added by the workshop organizers for Biocubes data. If the participant was not sure about the identification, she could turn on an “ID please flag” to encourage other iNaturalist community members to suggest one. After the information was input on this page, the participant clicked “save the observation,” then the data was added to iNaturalist. This data was linked to the Biocubes project page on iNaturalist and, consequently, the ant datum was officially shared publicly in an online environment for the first time. Anyone with Internet access can access this datum page for free. iNaturalist becomes the home repository for this Biocubes datum.

#### 5.4.2 Share to aggregator repository

That the Biocubes data were successfully shared on iNaturalist represents the data officially started their life cycle in the online environment. Being shared on iNaturalist is not the end point of their life, but the starting point for their future journey of being shared and used in a wider range of online and offline environments. This section focuses on investigating how the Biocubes data were shared to other online environments.

The next step in the online journey of the Biocubes data is the aggregator repository EOL. The data sharing contexts and processes on EOL had been investigated in-depth and reported in Chapter 4 so this dissertation could take advantage of studying each visible and invisible step necessary for data to be shared from their home repository (i.e., iNaturalist) to a larger online environment (i.e., EOL).

After analyzing the artifacts (e.g., iNaturalist and EOL webpages) and documentation data (e.g., EOL JIRA system content), three prerequisites must be met in order to share the Biocubes data from iNaturalist to EOL:

- Setting up data sharing partnerships;
- Creating shareable data by choosing an appropriate license; and
- Creating shareable data by assuring data quality.

#### 5.4.2.1 [Setting up data sharing partnerships](#)

The first precondition of sharing iNaturalist data with EOL is to set up an authoritative sharing channel by building an official partnership between iNaturalist and EOL.

Administrators and technical experts from both sides worked together for five months in 2012 to build this partnership. The major components of the collaborative efforts for building this partnership included:

- Reaching a mutual agreement to share media elements (i.e., photographs) of the data that reach research-grade;
- Registering a content partner account on the content partner management system for iNaturalist in order to document the data flow;

- Formulating data sharing strategies that satisfy both parties' needs;
- Creating user interface design solutions to allow data creators to easily select a license for their photographs;
- Preparing data files based on professional data transfer schema;
- Setting up a data transfer connector;
- Filtering out photographs without Creative Commons or Public Domain licenses and duplication of data provided by other EOL data partners;
- Previewing data sharing results (i.e., how iNaturalist data would look when they are displayed on EOL);
- Officially publishing iNaturalist data on EOL;
- Making the first versions of data updating plans for regularly re-harvesting iNaturalist data and updating new information added to previously shared data.

After the partnership had been successfully formed, the photographs of research-grade data with Creative Commons or Public Domain licenses were shared with EOL and displayed on the corresponding organism pages on EOL. Through March 2015, iNaturalist had shared 422,751 photographs with EOL. Without close collaboration between administrators and technical experts from both sides, transferring this large amount of data would be impossible.

#### 5.4.2.2 [Creating shareable data: choosing an appropriate license](#)

Although all observation data uploaded and displayed on iNaturalist can be accessed by anybody, it does not mean all the elements of the data can be shared with and used by others barrier-free. For any given datum on iNaturalist, the elements include:

- Organism name (taxon information)
- Observation date and time
- Observation location
- Observation description
- Data quality assessment results
- Community interaction with the data
- Media data (e.g., images, audio)

Among these elements, barriers appear when anyone considers sharing and using media data because they are considered intellectual property of individuals and are protected by certain types of copyright license. Therefore, those wishing to be the data users of this data must respect and follow the requirements of its copyright license. The options for copyright licenses for media data on iNaturalist include:

- All Rights Reserved
- Creative Commons
- Public Domain

In order to ensure the Biocubes data can be shared with EOL, the media data must adopt Creative Commons or Public Domain. Therefore, if Biocubes data contributors, including training workshop participants, would like their data be shared with EOL, they

must adopt one of these two licenses so that the media element of the data can become shareable.

### 5.4.2.3 Creating shareable data: Assuring data quality

Whether a datum is research-grade or not is calculated by iNaturalist's automatic data quality assessment algorithm. The results of the calculation for each observation datum are presented on the data quality assessment panel (Figure 5.8). According to the latest iNaturalist assessment criteria updated in August 2015, to reach research-grade a datum must include:

- A community-supported identification with a taxon level lower than family
- A digital voucher (i.e., the media data element, such as photograph or audio)
- A plausible observation time and location
- No disagreement from iNaturalist community members

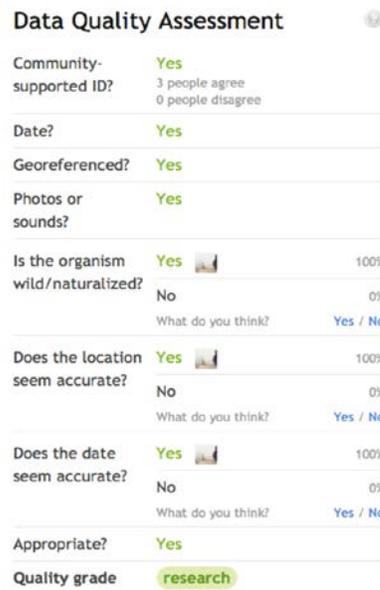


Figure 5.8 Example of the data quality assessment panel for a research-grade datum on iNaturalist.

In order to obtain a community supported identification, the data should be verified after each observation was collected and uploaded to iNaturalist. There are two formal ways to verify the data: 1) iNaturalist community members agree with the organism IDs provided by the observer and/or other iNaturalist community members; and, 2) iNaturalist community members suggest a new organism ID, usually a more specific label within the taxonomic hierarchy. A third optional way could be a helpful addition to the two formal ways: 3) iNaturalist community members leave comments asking for more details about the organism from the data creators/observers, such as the body features of an organism that were not clear or visible in the photographs, or the expected distribution of the same taxon group organisms.

For Biocubes data that were collected and uploaded in the training workshop, the project scientists (who were also iNaturalist community members) and voluntary iNaturalist community members reviewed the data according to their personal interests. They verified the accuracy of the data (e.g., identifications and location of organisms provided by data creators), improved the data quality (e.g., correcting a wrong identification or suggesting an identification with greater taxonomic resolution), or discussed with each other and the data creators in order to better identify the organisms. The data that had project scientists and/or iNaturalist community members' help with identification received a community-supported identification.

Up to March 2016, a total of 59 observation data points from the Biocubes workshop were uploaded to iNaturalist. However, although these data were observed during the

two days in the Biocubes training workshops (January 24–25, 2015), not all of them were uploaded during these two days: data were uploaded from January 26, 2015 to November 5, 2015. There could be two reasons for the delayed upload. The first could simply be that the project participants decided to upload the data after the workshop ended. The second reason is that the iNaturalist smartphone app did not successfully synchronize the data to the iNaturalist platform when the participants added the data to the iNaturalist app during the three-day workshop. The data could be finally have been automatically synchronized days, or even months, later when the iNaturalist app was updated, or the participants found the data did not upload to iNaturalist successfully, so then manually synchronized the data to ensure they were uploaded successfully. Table 5.3 shows how many data were uploaded to iNaturalist, by how many data creators and which Biocubes project.

Upload dates	The number of data points	The number of data provider(s)
November 5, 2015	9	1
June 29, 2015	1	1
January 30, 2015	4	1
January 27, 2015	1	1
January 26, 2015	11	2
January 25, 2015	9	5
January 24, 2015	24	5
Total	59	N/A

Table 5.3 Biocubes data collected in the Florida workshop by the data providers. The total number of data providers are not available because there are overlapping data providers for different days.

Among the 59 data points, only 13.56% (N = 8) of them became research-grade data.

Even though the percentage of research-grade data was not high, verifying data quality before sharing them to a wider range environments and audiences is necessary and

critical, otherwise data that are not accurate and/or contain errors will be widely distributed. In addition, among the 8 research-grade Biocubes data, three of them were changed from creative common license to all rights reserved license by a data provider after the workshop. Therefore, there were only 5 research-grade Biocubes data ultimately shared on EOL.

The photographs of research-grade Biocubes data were shared with EOL and displayed on the corresponding organism pages. Figure 5.9, Figure 5.10, and Figure 5.11 show an example of the photograph of a Biocubes research-grade datum shared with EOL.

← drop's observations

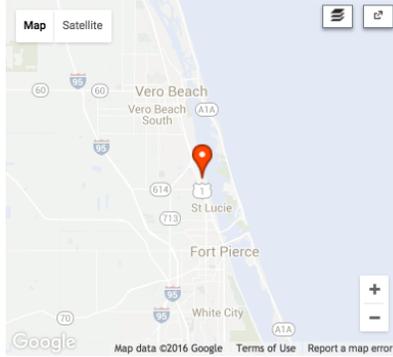
Previous Next

### Atlantic Ribbed Mussel (*Geukensia demissa*) · Observed by drop · January 24, 2015 · 04:40 PM EST

☆ Add to favorites Identify + Add to project → Share



Photo © drop, some rights reserved



Location: 3252-3270 Ocean Studies Dr,... (Google, OSM)  
Places: St. Lucie, Saint Lucie, US-FL, US Eastern States, US, North America, NA [More...](#)  
Lat 27.53411, Lon -80.355489  
Accuracy: 5m  
Geoprivacy: open [Hide details](#)

Added: Jan. 24, 2015 16:42:58 -0500

App: iNaturalist iPhone App

Number of Specimens:

Biocube ID: COSEE\_FL\_2015\_01

[Add/edit more fields](#)

drop's ID:  
**Atlantic Ribbed Mussel**  
(*Geukensia demissa*) [Agree?](#)

Community ID:  
**Atlantic Ribbed Mussel**  
(*Geukensia demissa*)

About

2 people agree

Identotron

#### Projects

Biocubes

View 10 from January 24, 2015 →

#### Comments & Identifications

- drop's ID: ~~Horse Mussel~~ (*Modiolus modiolus*) [Agree?](#)

Posted by drop over 1 year ago
- drop's ID: **Atlantic Ribbed Mussel** (*Geukensia demissa*) [Agree?](#)

Posted by [redacted] over 1 year ago (Flag)

Right family, wrong genus and species. This is not a horse mussel but a ribbed mussel. And if it were a horse mussel, in Florida the only horse mussel species are *Modiolus squamosus* and *Modiolus americanus*.

Ribbed mussels live in the mud of the back bay in the salt marshes; they are easy to find there and there are a lot of them. In contrast, horse mussels live on the outer coast, and subtidally; you don't find them that often.

In Florida there is also supposed to be a related species or subspecies *Geukensia granosissima*, which has finer sculpture on the shell.

Posted by [redacted] over 1 year ago
- Thanks for the help! We were looking around the wrack line near a rocky intertidal zone.

Posted by drop over 1 year ago
- drop's ID: **Atlantic Ribbed Mussel** (*Geukensia demissa*)

Posted by [redacted] over 1 year ago (Flag)
- Oh judging from the map I thought you were on the back bay. When you say rocky intertidal do you mean there was a rock jetty there or something?

These ribbed mussels sometimes get washed out of the mud habitat and can then end up washing up anywhere, still live, even on the outer coast.

Posted by invertzoo over 1 year ago
- drop's ID: ~~Atlantic Ribbed Mussel~~ (*Geukensia demissa*)

Posted by drop over 1 year ago
- drop's ID: **Atlantic Ribbed Mussel** (*Geukensia demissa*)

Posted by drop over 1 year ago

#### Data Quality Assessment

Community-supported ID?	Yes	2 people agree 0 people disagree
Community ID at species or lower?	Yes	
Community can confirm/improve ID?	No	What do you think? <a href="#">Yes</a> <a href="#">No</a>
Date?	Yes	
Georeferenced?	Yes	
Photos or sounds?	Yes	
Is the organism wild/naturalized?	Yes	100%
Does the location seem accurate?	Unknown	What do you think? <a href="#">Yes</a> / <a href="#">No</a>
Does the date seem accurate?	Unknown	What do you think? <a href="#">Yes</a> / <a href="#">No</a>
Appropriate?	Yes	Inappropriate? <a href="#">Flag this observation</a>
Quality grade	Research	

[Hide details](#)

#### External Links

This observation has been incorporated into the following external websites:

GBIF

Observation © drop

BY-NC some rights reserved

Figure 5.9 A Biocubes datum on iNaturalist (drop, 2015, January).

*Geukensia demissa*
add to a collection

Ribbed Mussel
[learn more about names for this taxon](#)

Overview
Detail
Data
17 Media
7 Maps
Names
Community
Resources
Literature
Updates



*Geukensia demissa* TRUSTED

 © drop  
Source: [iNaturalist.org](#)

[see all media](#)

**EOL has data for 23 traits** [see all](#)

<b>body length (CMO)</b>	(max) (measurement)	293 mm 45 mm
<b>shell length, bivalve</b>	(max)	115 mm
<b>shell height, bivalve</b>	(max)	293 mm
<b>first appearance (older bound)</b>		2.59 million years ago
<b>habitat</b>		anthropogenic geographic feature aquatic habitat bay <a href="#">more</a>
<b>preys upon</b>		<a href="#">Chione californiensis</a> <a href="#">Chione</a> <a href="#">Ilucifraga</a> <a href="#">Cirriformia</a> <a href="#">spiralbrancha</a> <a href="#">more</a>
<b>has predator</b>		<a href="#">Geukensia demissa</a>
<b>interacts with</b>		<a href="#">Geukensia demissa</a>
<b>interacts with</b>		<a href="#">Calidris alpina</a> <a href="#">Callinectes</a> <a href="#">sapidus</a> <a href="#">Dyspanopeus</a> <a href="#">sayi</a> <a href="#">more</a>

**Comprehensive Description** [read full entry](#)

[learn more about this article](#)

*Geukensia demissa* is a member of the family Mytilidae. The surface of the shell is grooved or ribbed and oval in shape. The ribbed mussel has a narrow blunt pointed head that is attached to submerged substrata. Shells are usually glossy appearing olive-brown to brown-black with some yellow to a white on the outside and white on the interior with purplish tints.

TRUSTED CC BY-NC-SA © Smithsonian Marine Station at Fort Pierce  
• Source: [Indian River Lagoon Species Inventory](#)

**Present in 38 collections** [see all](#)

-  [Salt Marsh Species](#)  
14 other items
-  [So Cal Pleist fauna](#)  
99 other items
-  [EOL Hotlist of High Priority Taxa](#)  
91958 other items

This taxon hasn't been featured in any communities yet.

[Learn more about Communities](#)

**Found in 28 classifications** [see all](#)

Species recognized by [NCBI Taxonomy](#):

- [Cellular organisms](#) +
- [Eukaryota](#) +
- [Opisthokonta](#) +
- [Metazoa](#) +
- [Eumetazoa](#) +
- [Bilateria](#) +
- [Protostomia](#) +
- [Lophotrochozoa](#) +
- [Mollusca](#) +
- [Bivalvia](#) +
- [Pteriomorpha](#) +
- [Mytiloidea](#) +
- [Mytilidae](#) +
- [Modiolinae](#) +
- [Geukensia](#) +
- [Geukensia demissa](#)**
- [Geukensia granosissima](#)

Figure 5.10 The Biocubes datum (i.e., photograph) is shared from iNaturalist to EOL platform (Ribbed Mussel, n.d.).

Image of *Geukensia demissa*  
© drop add to a collection



[View full-size image](#)

**In the latest image**

[Geukensia demissa](#) TRUSTED

---

**Source information**

 BY-NC

Copyright drop, licensed under a Attribution-NonCommercial License license: <http://creativecommons.org/licenses/by-nc/3.0/>  
© drop  
[View source](#)  
Supplier: [iNaturalist.org](http://iNaturalist.org)  
Publisher: inaturalist  
Creator: drop  
Location: 3252-3270 Ocean Studies Dr, Fort Pierce, FL, US  
[View full-size image](#)

**Image rating** [learn about rating](#)

5 stars	<input type="checkbox"/>	0
4 stars	<input type="checkbox"/>	0
3 stars	<input type="checkbox"/>	0
2 stars	<input type="checkbox"/>	0
1 star	<input type="checkbox"/>	0
default rating	★ ★ ★ ★ ★	
your rating	★ ★ ★ ★ ★	

---

**Revisions**

2015-03-15 03:00:43 UTC

**Latest updates**

No one has provided updates yet.

[Learn how to contribute.](#)

Figure 5.11 The Biocubes datum (i.e., photograph) is shared from iNaturalist to EOL platform, with detailed data source information (drop, 2015, March).

## 5.5 Conclusion

The findings in this chapter are summarized regarding the three research questions as follows:

- Who are the data providers in the CyberSEES project?

The individual-level non-professional data providers observed in a citizen science project training workshop are the educators from both formal and non-formal education institutions.

- Who are the data sharing mediators?

There are three groups of *human* mediators in this case. The first group of human mediators are the Biocubes project organizers (i.e., the CyberSEES project

members). The second group of human mediators are the directors of iNaturalist, the members from EOL SPG, and an EOL technician. The third group of human mediators are the iNaturalist community members.

The *technology* mediators are embedded in the iNaturalist technology infrastructure, including three information systems: the data entry system, the organism observation page, and the data export system.

- What are the data sharing processes?

Three general steps of the processes are identified in this chapter.

- With guidance provided by the citizen science project organizers, the individual-level data providers generated biodiversity data in offline environment and then shared the data in an online environment—the home repository of the data—via using a smartphone app;
- The human mediators from the home repository and the aggregator repository set up authoritative data sharing partnerships;
- Provided the data gain appropriate licenses and reach a certain quality level (i.e., research-grade), the data is aggregated by the aggregator repository via the established data connector between the home repository and the aggregator repository.

## 6 Discussion

### 6.1 Overview of the chapter

In order to answer the overarching research question: *how data are shared effectively across research and public communities*, this dissertation investigated the data sharing practices in two cases, EOL and CyberSEES. EOL is a large-scale aggregator repository, whereas CyberSEES is a cyberinfrastructure project. The two cases share the same goal of facilitating data sharing in both research and public communities. Three research sub questions were asked in each case: who are the data providers, who are the mediators, and what are the data sharing processes. The answers to these three questions constitute a comprehensive understanding of the data sharing practices of sharing data effectively across research and public communities.

This chapter includes four sections. The first compares the findings of the two cases. The second section links the findings back to the theoretical framework and develops a new integrated theoretical framework of data sharing across communities. The last two sections discuss the implications for data sharing practices and data sharing infrastructure design.

### 6.2 Summarizing and comparing two cases and their findings regarding the research questions

The two cases are different from each other in multiple aspects. In terms of data sharing contexts, they support different levels of sharing by different level data providers, provide different scales of online environments, and adopt different sharing platforms. With

respect to data sharing processes, they each developed and designed different ways of sharing that match different levels of data sharing (i.e., organizational level and individual level). These two cases show the multi-level nature of data sharing phenomena. Table 6.1 is adapted from Table 3.1, summarizing the differences between the two cases.

<b>Cases</b>		<b>Encyclopedia of Life (EOL)</b>	<b>The CyberSEES project (CyberSEES)</b>
<b>Different aspects</b>			
<b>(Online) data sharing contexts</b>	<b>Different levels of data providers</b>	For collective-level data providers	For individual-level data providers
	<b>Different scales of online environments</b>	Aggregating and presenting different types of biodiversity data that constitute the knowledge about life on Earth	Collecting and presenting biodiversity observation/occurrence data
	<b>Sharing platforms</b>	EOL	iNaturalist/Biocubes project page
<b>Data sharing processes</b>		Collective-level data providers build formal data sharing partnerships with EOL to transfer their data on EOL platform.  (i.e., formal collaboration relationship between data providers and EOL)	Individual-level data providers participate in the citizen science project called Biocubes, collect data in the fields, and share data on iNaturalist platform.  (i.e., informal collaboration between the citizen science project participants and citizen science project organizers/scientists)
		<ul style="list-style-type: none"> <li>• iNaturalist as a collective-level data provider builds formal data sharing partnership with EOL.</li> <li>• iNaturalist community members helped to improve the data quality to research-grade quality.</li> <li>• The Biocubes research-grade data on iNaturalist platform are shared with EOL platform.</li> </ul>	

Table 6.1 The differences between the two cases from the perspective of data sharing contexts and data sharing processes.

The different aspects summarized in Table 6.1 affect and interact with each other. EOL provides a large-scale online environment for aggregating and sharing the data via its own powerful infrastructure and tools. They formalized the processes of sharing data to ensure the reliability and trustworthiness of the data sources and the data sharing channel. CyberSEES used a citizen science project called Biocubes as a vehicle for collecting citizen science data and studying the infrastructure development and design for sharing citizen science data. CyberSEES and Biocubes provided a relatively small-scale online environment by adopting the infrastructure and tools built for the iNaturalist community. As an online community of naturalists, iNaturalist encourages any individuals and projects (e.g., citizen science projects) to share biodiversity observation data on their platform. EOL aggregates biodiversity data collected by both researchers and non-professionals, provided the data created by non-professionals reaches a certain level of quality. Biocubes focused on recruiting non-professionals and encouraging them to create biodiversity data, and facilitating the data to be shared in online environments.

These different aspects regarding online data sharing contexts and processes are helpful for selecting the theoretical representative cases in this dissertation. However, they far from describe all the differences between the two cases, a description far beyond the scope of the discussion in this dissertation. There are, however, a few deeper core differences should be acknowledged for their important impact on the findings revealed in the two cases.

The deeper differences between the two cases are rooted in the differences between their knowledge infrastructures (Edwards, 2010). In other words, the people, artifacts, and institutions networked by EOL and CyberSEES are different, as are how they were formed and how they function. Some of these major differences are reflected by the development history for each of these cases, for example the age of the knowledge infrastructure, individual and institutional support, and the funding sources that make the development of the knowledge infrastructure possible. The visible parts of the knowledge infrastructure in the two cases became available to the general public in an online environment when they were formally launched; this is the point at which this dissertation started to collect data to understand their data sharing practices. Table 6.2 summarizes the major differences between the knowledge infrastructures of EOL and CyberSEES.

Differences of the knowledge infrastructures	EOL	The CyberSEES Project	
		The Biocubes project	iNaturalist
Age (years)	8+ (officially launched in February 26, 2008)	1+ (officially launched in January 1, 2015)	8+ (officially launched in 2008)
Individual support	Dr. Edward O. Wilson	<ul style="list-style-type: none"> <li>Individual researchers from Smithsonian institution and University of Maryland</li> <li>A freelance photographer David Liittschwager</li> </ul>	<ul style="list-style-type: none"> <li>Three master students: Nate Agrin, Jessica Kline, Ken-ichi Ueda</li> <li>Individual developer: Sean McGregor</li> <li>Collaborator: Scott Loarie</li> </ul>
Institutional support	<ul style="list-style-type: none"> <li>Field Museum</li> <li>Harvard University</li> <li>The Marine Biological Laboratory</li> <li>Missouri Botanical Garden</li> <li>Smithsonian Institution</li> </ul>	<ul style="list-style-type: none"> <li>Smithsonian Institution</li> <li>University of Maryland</li> </ul>	<ul style="list-style-type: none"> <li>California Academy of Sciences (since 2014)</li> </ul>
Funding sources	<ul style="list-style-type: none"> <li>Philanthropic nonprofit organization</li> <li>Private independent grantmaking institution</li> <li>Other research and education institutions</li> </ul>	<ul style="list-style-type: none"> <li>Federal agency</li> </ul>	<ul style="list-style-type: none"> <li>Scientific and educational institution (since 2014)</li> </ul>
The amount of funding	\$12.5 million (Seed funding)	\$371,045	Unknown
End date	N/A	December 31, 2016	N/A

Table 6.2 Summary of the major differences between the knowledge infrastructures of the two cases.

In the case of EOL, it started in 2007 with Dr. Edward O. Wilson's TED Prize Speech and five original cornerstone institutions (i.e., Field Museum, Harvard University, the Marine Biological Laboratory, Missouri Botanical Garden, and the Smithsonian Institution). Two foundations, the American philanthropic nonprofit organization, Alfred P. Sloan Foundation, and a US private independent grantmaking institution, John D. and Catherine T. MacArthur Foundation, provided generous seed funding (\$2.5 million and \$10 million respectively) to build the unlimited online encyclopedia of all named species on Earth (EOL History, n.d.; Alfred P. Sloan Foundation, 2016). Additional funding is from the five cornerstone institutions. After EOL was officially launched in 2008, continuing funding is used to further develop EOL. There is no set end date for EOL.

CyberSEES officially started in 2015. The individual researchers from two organizations, the Smithsonian Institution and University of Maryland, proposed and developed this project with the consent and support from freelance photographer David Liittschwager who invented and built biocubes for use in his work. Having received this consent, researchers could develop Biocubes as a citizen science project. CyberSEES NSF funding was distributed to the Smithsonian Institution and University of Maryland separately, receiving \$271,045 and \$100,000 respectively (McKeon & Meyer, 2015; Wiggins & Preece, 2015). Only part of this funding was used for developing Biocubes. CyberSEES' estimated end date is December 31, 2016.

Since iNaturalist played an important role in the knowledge infrastructure of the CyberSEES project, it is worth to mentioning its history, although it is completely

independent from Biocubes or CyberSEES. iNaturalist was started as the Master's final project of three graduate students, Nate Agrin, Jessica Kline, and Ken-ichi Ueda, at UC Berkeley's School of Information in 2008 and aimed to help individual naturalists to share their observations online (Loarie, 2016a). Independent developer, Sean McGregor, continued working on developing the iNaturalist site and, in 2011, Ueda collaborated with Scott Loarie to evolve the site into a Limited Liability Company and significantly expanded its organizational-level collaboration network with other platforms and organizations (Loarie, 2016a). In 2014, it was acquired by scientific and educational institution, California Academy of Sciences (CAS); since then, iNaturalist serves as CAS's online social network for naturalists (Loarie, 2016a). Like EOL, there is no set end date.

As CyberSEES will end at the end of 2016, uploading Biocubes data to the iNaturalist platform is an effective strategy to prolong the life of the Biocubes data. Due to the formal data partnership between EOL and iNaturalist, research-grade Biocubes data are further shared with EOL, ensuring high quality Biocubes data are not only given a long life, but also endowed with more important meaning and purpose. On both iNaturalist and EOL, Biocubes data are shared successfully across research and public communities, which would not have been possible if EOL, CyberSEES, Biocubes, and iNaturalist do not share the core mission of collecting data not only created by researchers but also non-professionals and sharing that data openly in online environments for the benefit of both research and public communities. All the efforts made by the human workers of EOL, CyberSEES, Biocubes, and iNaturalist further this mission.

Besides being linked by this mission, EOL, CyberSEES, Biocubes, and iNaturalist are also linked by the formal data sharing partnership established between EOL and iNaturalist and by research-grade Biocubes data aggregated by EOL. The core part of the work to build the data sharing partnership was to establish a technology data sharing connector between EOL and iNaturalist. This connector bridges the EOL online environment to the iNaturalist online environment so that research-grade data on iNaturalist can be transferred from there to EOL. As part of iNaturalist, research-grade Biocubes data is transferred to EOL. It is this research-grade Biocubes data that connects EOL and the Biocubes project—without research-grade data, EOL and Biocubes cannot be linked in an online environment, even if a partnership between EOL and iNaturalist preexisted.

Compared with the large-scale data on EOL (i.e., its coverage of different types of data for all life on Earth), Biocubes is a very small-scale data source. However small, this data source still requires CyberSEES to provide it an appropriate social situation needed for collaboratively creating and sharing the data created by non-professionals. Compared with researchers who create their own social situations within which to collect data, non-professional data creators require a social situation to be prepared and built for them so as to conduct the activities of collecting and sharing data.

The data sharing journey from Biocubes to EOL illustrates the comprehensive cross-level processes from the origination of the data created by non-professionals in the offline environment to being widely shared in the same large-scale online environment as data

created by researchers. Therefore, the two cases provide a unique research opportunity for understanding the detailed data sharing practices in this data sharing journey. This dissertation took this research opportunity and investigated these data sharing practices by focusing on answering three research questions:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

Chapters 4 and 5 reported the findings of each case separately. This section brings the findings from the two cases together to compare them with respect to each research question; Table 6.3 summarizes this comparison.

Research questions	EOL	CyberSEES
<b>Q1: Who are the Providers?</b>		
	<p><b>Collective level:</b> 329 content partners</p> <ul style="list-style-type: none"> <li>• Venerable organizations (N= 20)</li> <li>• Professional repositories (N = 148)</li> <li>• Citizen science initiatives (N = 5)</li> <li>• Social media platforms (N = 7)</li> <li>• Education communities (N = 13)</li> <li>• Subsidiaries (N = 95)</li> <li>• Academic papers (N = 20)</li> <li>• Other (N = 21)</li> </ul>	<p><b>Individual level:</b></p> <ul style="list-style-type: none"> <li>• Biocubes project participants (i.e., educators) from education community (N =13)</li> <li>• Project organizers (i.e., biologists, educators, social scientists) from the research community and education community (N = 11)</li> </ul> <p><b>Collective level:</b></p> <ul style="list-style-type: none"> <li>• The Biocubes project</li> <li>• The iNaturalist community/repository</li> </ul>
<b>Q2: Who are the mediators?</b>		
Core human mediators	<p><b>Collective-level sharing:</b> sharing data from content partners to EOL</p> <p><b>EOL side:</b></p> <ul style="list-style-type: none"> <li>• Members from EOL working groups: Species Page Group (SPG) and Biodiversity Informatics Working Group (BIG)</li> <li>• EOL contractor developer</li> </ul> <p><b>Content partner side:</b></p> <ul style="list-style-type: none"> <li>• Data managers</li> <li>• Data technicians</li> <li>• Data contributors</li> </ul>	<p><b>Individual-level sharing:</b> sharing data from an offline environment to iNaturalist (i.e., home repository)</p> <ul style="list-style-type: none"> <li>• Biocubes project organizers</li> </ul> <p><b>Collective-level sharing:</b> sharing data from iNaturalist to EOL (i.e., aggregator repository)</p> <ul style="list-style-type: none"> <li>• Members from EOL working groups and EOL contractor developer and iNaturalist directors</li> <li>• iNaturalist community members</li> </ul>
Technology mediators		
<i>For data providers</i>	<ul style="list-style-type: none"> <li>• Content partner management system</li> <li>• EOL Species pages</li> </ul>	<ul style="list-style-type: none"> <li>• iNaturalist data entry system</li> <li>• iNaturalist organism page</li> </ul>
<i>For data users</i>	<ul style="list-style-type: none"> <li>• EOL Species pages</li> <li>• EOL data export tool: TraitBank</li> </ul>	<ul style="list-style-type: none"> <li>• iNaturalist organism page</li> <li>• iNaturalist data export tool</li> </ul>
<b>Q3. What are the data sharing Processes?</b>		
	<ul style="list-style-type: none"> <li>• Preparing social relationships and reaching mutual agreement</li> <li>• Developing a data sharing connector</li> <li>• Updating data and/or data sharing connector</li> </ul>	<ul style="list-style-type: none"> <li>• Generating data in offline and online environment for the first time</li> <li>• Turing data into shareable data</li> <li>• Setting up data sharing partnerships</li> </ul>

Table 6.3 The comparison of findings between case one (EOL) and case two (CyberSEES).

### 6.2.1 Data providers: different types of data providers at different levels

The data providers are an individual or group of human actors who: 1) have ownership of or management rights to the data; 2) have the willingness to share their data publicly; and, 3) take action to share the data in one or multiple online environment(s). Table 6.4 summarizes the different levels of data provider in EOL and CyberSEES identified in this dissertation.

Case name	Q1: Who are the Providers?	
	EOL	CyberSEES
<b>Individual-level data providers</b>	N/A	<ul style="list-style-type: none"> <li>• Biocubes project participants (i.e., educators) from education community (N = 13)</li> <li>• Project organizers (i.e., biologists, educators, social scientists) from the research community and education community (N = 11)</li> </ul>
<b>Collective-level data providers</b>	329 content partners <ul style="list-style-type: none"> <li>• Venerable organizations (N= 20)</li> <li>• Professional repositories (N = 148)</li> <li>• Citizen science initiatives (N = 5)</li> <li>• Social media platforms (N = 7)</li> <li>• Education communities (N = 13)</li> <li>• Subsidiaries (N = 95)</li> <li>• Academic papers (N = 20)</li> <li>• Other (N = 21)</li> </ul>	<ul style="list-style-type: none"> <li>• The Biocubes project</li> <li>• The iNaturalist community/repository</li> </ul>

Table 6.4 Different level of data providers in the case of EOL and the case of the CyberSEES project.

EOL’s data providers, or “content partners,” built formal data sharing partnerships with EOL in order to share their data on the EOL platform. These data providers adopted

collective-level identities when they introduced themselves in online environments. This dissertation categorized these identities into seven types, representing seven different types of collective-level data providers: venerable organizations, professional repositories, citizen science initiatives, social media platforms, education communities, subsidiaries, and academic papers. Behind the collective-level identities of many content partners were a limited number of human actors who worked on managing the large-scale data. Sharing this data with EOL or other platforms is part of their data management practices.

Venerable organization type data providers have more human actors involved in building their partnership with EOL than other types of collective-level data providers.

Organization type data providers are more likely to have more complex organizational structures than other types of data providers, such that a single person could not make a decision (i.e., a data sharing decision) on behalf of the entire organization. Therefore, this type of data provider needs to involve more human mediators than other types.

More specifically, although these data providers adopt the identity of an organization, the organizations as entities are not shared with EOL but rather, it is their data products (i.e., database, repository) that are. When an administrative actor of an organization (e.g., the director) agrees to share their data product with EOL, the data sharing processes start and s/he needs to direct EOL to the administrative actors responsible for producing and managing this data product (e.g., informatics director or manager). The administrative

actors of the data product might direct EOL to one or more executive actors (e.g., developers) in order to build the data sharing connectors.

Many of EOL's data providers only speak for the data product itself, which could explain why many professional repository type data providers are hosted by organizations but did not adopt the identity of the organization. These providers are allowed to be relatively independent from the organizations who host them and therefore might not need to involve as many human mediators as organization type data providers when working with EOL build data sharing partnerships. For example, typical professional repositories utilized only two human mediators for building the collaborative partnership with EOL; one of them played an administrator role and the other a technician role. It was not uncommon for only one human mediator to play both roles.

Other types of collective-level data providers, such as citizen science initiatives, social media platforms, and education communities, are more likely to have organizational structures similar to professional repository data providers. Therefore, only a few human mediators were utilized in building the data sharing partnership with EOL. However, as professional repositories, citizen science initiatives, and social media platforms grow, their organizational structures could become increasingly complex, increasing the number of human mediators that may need to be involved in future sharing. This accounts for the EOL JIRA system content that showed one professional repository data provider involving more human mediators than any other: the organizational structure of this professional repository is similar to a venerable organization data provider.

The data providers identified in the case of EOL are all at the collective-level. In contrast, both individual-level data providers and collective-level data providers are identified in CyberSEES/Biocubes. There are two groups of individual-level data providers: the first includes Biocubes project participants (i.e., the individual educators from the education community) who collected and shared data on iNaturalist; the second group includes Biocubes organizers who are also CyberSEES project members from both the research and education communities. Two collective-level data providers are identified in CyberSEES: the first is the Biocubes project itself as a citizen science project and data source and the second is iNaturalist as an online community of naturalists and a citizen science repository.

The individual-level data providers hold the smallest amount of data and while they collected the data by themselves or collaboratively with others, they did not necessarily submit the data to iNaturalist by themselves. They have the ownership of the data so, should they choose to submit it to iNaturalist by themselves, they could freely choose the copyright for the entire observation data and each specific element (i.e., media) of it. They can always come back to edit the data, including changing the copyright after they created the data. They have the highest power of control over their data. However, if they do not upload the data using their iNaturalist account, they lose the control of the data unless they contact the person who uploads the data for them.

Collective-level data providers hold more data than the individual-level data providers. For example, the Biocubes project holds all Biocubes data collected by individual-level data providers. iNaturalist not only holds all Biocubes data uploaded and shared on iNaturalist, but also holds all the other data uploaded and shared on its platform by individual-level data providers. However, Biocubes and iNaturalist does not truly have ownership of these data, but rather just the rights to manage it.

Collective-level data providers are granted permission to further share individual-level data providers' data if these providers choose a Creative Commons or public domain license. Therefore, although collective-level providers hold much more data than individual-level providers, if they do not have the ownership of the data or the data do not have appropriate licenses, they cannot share or re-use these data. In this case, collective-level data providers might not be considered data providers, because they could not share the data they possess, although they can always attempt to obtain permission from the individual-level data provider. However, when data are created by many non-professionals (e.g., crowd data), it is not realistic to ask permission from each individual-level data provider. Therefore, it is important to promote an open sharing culture not only in the research community, but also in the public community.

In addition, the relationship between the two collective-level data providers, Biocubes and iNaturalist, shows a multi-layer data sharing relationship. In the first layer, Biocubes shares data with iNaturalist.; in the second, iNaturalist shares data with EOL. The third and final layer is Biocubes data being shared with EOL: although Biocubes provided data

to EOL via iNaturalist, it did not build a formal data sharing partnership with EOL directly. This does not take away from the fact that its data are shared on EOL by taking advantage of the existing formal partnership between iNaturalist and EOL. However, it is also restricted by these multiple layers. iNaturalist and EOL data sharing partnership allows the photograph element of research-grade data to be transferred to and shared on EOL, but not the entire observation data (i.e., all elements of the data). This means that only the photographic element of Biocubes data can be shared with EOL; all the other Biocubes data elements are left unshared.

### 6.2.2 Human mediators: different structures

Human mediators are any human actors who work on facilitating data sharing from creators to users. They are the indispensable parts that comprise the human infrastructure of the entire knowledge infrastructure. This dissertation focused on the core human mediators who directly work with the data providers, “core” being determined by whether data sharing can be carried out without their effort. Table 6.5 shows the summary of the human mediators in the case of EOL and the case of the Biocubes project.

Case name	Q2: Who are the mediators?	
	EOL	CyberSEES
<b>Human mediators</b>	<ul style="list-style-type: none"> <li>• EOL working group members from SPG and BIG, EOL contractor developers</li> <li>• Content partners’ data managers, data technicians, and data contributors</li> </ul>	<ul style="list-style-type: none"> <li>• Biocubes project organizers</li> <li>• iNaturalist community members</li> <li>• EOL working group members, EOL contractor developer, and iNaturalist directors</li> </ul>

Table 6.5 The human mediators in EOL and Biocubes.

In EOL, the human mediators are those who directly worked on building the partnerships with content partners and exist on both the EOL and content partner side. The core human mediators in EOL include the members from the SPG and BIG, as well as a contractor developer. Core human mediators for the content partners could be data managers, data technicians, and/or data contributors. Since the EOL core human mediators were not changed no matter the content partner or changes are made by the content partner side, they are the joint point that connects all EOL content partners. Furthermore, regardless of the relationships between different content partners outside of EOL, they are equal and independent to one another on EOL.

In the case of CyberSEES, there are three layers of human mediators. The first includes Biocubes project organizers who are also CyberSEES project members. Second-layer human mediators include EOL working group members and iNaturalist data managers. Third-layer human mediators include iNaturalist community members. At the macro-level, the first-layer human mediators introduce outsiders to the research community (i.e., non-professionals from the general public) into the research community and train them to become research data creators (e.g., citizen scientists) who created (potentially) shareable research data. At the micro-level, the first-layer human mediators introduce new users to the iNaturalist community and train them to become iNaturalist users (at least one-time users). The second-layer mediators build the partnerships with data partners, such as EOL, by reaching a collaboration agreement and setting up the technology data connector with EOL. Finally, the third-level human mediators include iNaturalist community

members who help to turn potentially shareable research data to shareable research data (i.e., research-grade data). Although these three layers are located in the same online environment, iNaturalist, they are relatively independent of one another and all have a different focus regarding facilitating data sharing.

The findings about human mediators in both cases illustrates different structures human mediators can take. Figure 6.1 and Figure 6.2 show two fuzzy models of human mediator structures for EOL and Biocubes respectively and can be applied more broadly to two types of repository: aggregator repository and home repository respectively. In Figure 6.1, the circle in the center represents EOL human mediators; the surrounding circles represent different content partner's human mediators. In Figure 6.2, the ellipse represents the iNaturalist online environment and three rectangles inside it represent different layers of human mediators: first-layer human mediators facilitate data sharing from an offline to an online environment (i.e., iNaturalist); second-layer human mediators facilitate data sharing from iNaturalist (i.e., home repository) to another online environment (e.g., aggregator repository, EOL); and third-layer human mediators improve the data introduced in iNaturalist by the first-layer human mediators, so that the data can in fact be shared with other online environments via the data sharing connector built by the second-layer mediators. Figure 6.2 reveals the details of what is happening within one of the surrounding circles in Figure 6.1.

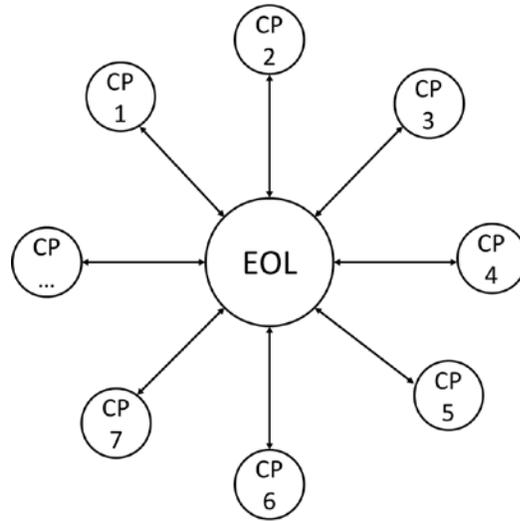


Figure 6.1 Human mediators structure on EOL, an aggregator repository.

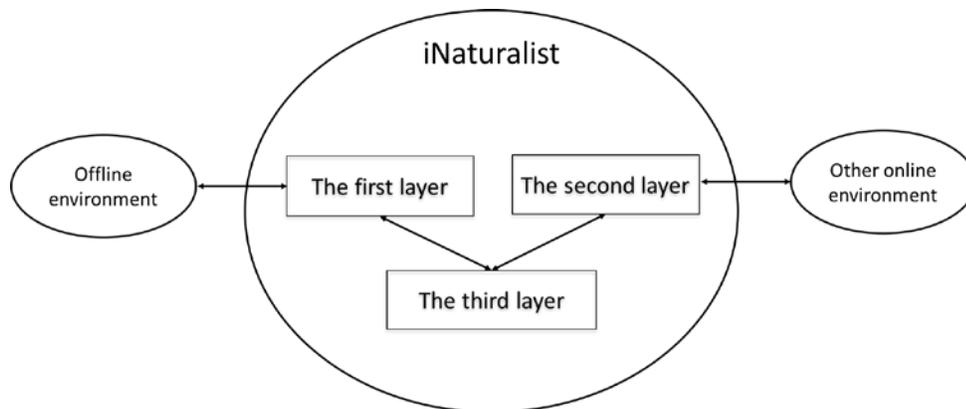


Figure 6.2 Human mediator structure on iNaturalist, a home repository.

### 6.2.3 Technology mediators: barriers vs. no barriers

Technology mediators are designed and developed for satisfying the technology demands of connecting data providers and users in various ways. Technology mediators are essential parts of the technology infrastructure within the entire knowledge infrastructure and are influenced by the design and development of the technology infrastructure.

These mediators support connecting the data providers to the technology infrastructure so

that the providers can share the data via the technology mediators. Technology mediators also support connecting data users to the technology infrastructure so they can access data providers' data. Table 6.6 shows the summary of the technology mediators in EOL and Biocubes.

Case names		Q2: Who are the mediators?	
		EOL	CyberSEES
<b>Technology mediators</b>	For data providers	<ul style="list-style-type: none"> <li>• Content partner management system</li> <li>• EOL Species pages</li> </ul>	<ul style="list-style-type: none"> <li>• iNaturalist data entry system</li> <li>• iNaturalist organism page</li> </ul>
	For data users	<ul style="list-style-type: none"> <li>• EOL Species pages</li> <li>• EOL data export tool: TraitBank</li> <li>• EOL API</li> </ul>	<ul style="list-style-type: none"> <li>• iNaturalist organism page</li> <li>• iNaturalist data export tool</li> <li>• iNaturalist API</li> </ul>

Table 6.6 The technology mediators in EOL and CyberSEES.

### 6.2.3.1 Connecting the data providers to the technological infrastructure

EOL's content partner management system—and data entry system—is embedded in EOL's user account system. A content partner must become a regular registered EOL user, in other words an EOL community member, prior to becoming a content partner. Therefore, a content partner has a two-layer account in the EOL infrastructure: a regular user account and a content partner account. A content partner needs to have a content partner account in order to upload data source files to EOL and to have its name and information automatically displayed on the EOL website.

Merely creating a content partner account does not mean a real content partner is created: it is still pending. SPGers need to review the data source that the data providers plan to share for reliability and trustworthiness, after which the human mediators from both EOL and the content partner proceed to the next step of building the partnership. If a data

provider wants to share their data on EOL but does not want to or is not able to build a formal partnership with EOL, they can first share their data with an existing content partner. This indirect way of sharing data on EOL is effective, as was shown by CyberSEES who share Biocubes data on EOL in this way by sharing the data to iNaturalist, an EOL content partner.

However, not all EOL content partners created their content partner accounts themselves. Some content partners provide the data source files to EOL, but do not have time or feel they need to manage the content partner management system themselves. Therefore, human mediators need to create content partner accounts for them having asked them to provide the information such as a log, a short paragraph introduction, and so on to add to their accounts. Having completed this, EOL human mediators would email the content partners their account information and a brief guide about how to use it.

In the CyberSEES project and similar to EOL, before sharing Biocubes data, a data provider must become a registered iNaturalist user and therefore become a iNaturalist community member. Like EOL, the data entry system is embedded in the user account system. However, unlike EOL's need for human mediators to review the data source, anybody can directly upload biodiversity data on iNaturalist without being reviewed by iNaturalist staff.

An additional step is needed for connecting Biocubes data providers to the right place: the Biocubes project organizers needed to create a project page on iNaturalist. Once one

of the Biocubes project organizers created a regular iNaturalist user account, s/he created the Biocubes project page which is embedded in his/her regular iNaturalist user account. The project information does not need to be reviewed by iNaturalist staff before it is published.

#### 6.2.3.2 Connecting the data users to the technology infrastructure

There are three ways that data users can be connected with the data and the data providers. Firstly, data users access the data through browsing web pages containing the data. Secondly, data users use the data export tool to directly download the data. Lastly, data users can use a web service (e.g., API) to download the data. In both EOL and CyberSEES, all three ways are available to any data user. Presenting data on the web pages and providing a direct dataset download function are important for individual data users, while providing API could be critical for some collective-level data users, especially aggregator-level data users.

EOL's species pages contain comprehensive information (i.e., various forms of data) about a specific organism provided by diverse data providers (i.e., content partners). In the case of Biocubes, the organism observation pages on iNaturalist only include basic information about a specific organism (i.e., metadata) provided by a single observer account (although there might be a group of people behind a single observer account). Whereas on EOL there is only one species page for a single organism, on iNaturalist there could be many organism observation pages. These observation pages could be

generated by different observers and show the occurrence of this specific organism at the same or different times and places.

EOL’s data export tool now allows data users to use the search filter to choose the taxon and attributes of organisms. The iNaturalist data export tools support more filter choices, including taxon, observation date, time, location, the observer, the observation licenses, and the project the data are linked to.

#### 6.2.4 Data sharing processes

Most previous studies about data sharing focused on facilitating researchers sharing data with other researchers. The dissertation moves two steps forward: to facilitate sharing data with not only researchers, but also non-professionals, and to facilitate sharing data created by non-professionals in the same ways as that created by researchers. Therefore, the data sharing processes revealed in this dissertation concern sharing data created by non-professionals across research and public communities. Table 6.7 shows the summary of the general steps included in the data sharing processes.

<b>Q3. What are the data sharing Processes?</b>		
<b>Case name</b>	<b>EOL</b>	<b>CyberSEES</b>
<b>Data sharing process</b>	<ul style="list-style-type: none"> <li>• Preparing social relationships and reaching mutual agreement</li> <li>• Developing a data sharing connector</li> <li>• Updating data and/or data sharing connector</li> </ul>	<ul style="list-style-type: none"> <li>• Generating data in offline and online environment for the first time</li> <li>• Setting up data sharing partnerships</li> <li>• Turing data into shareable data</li> </ul>

Table 6.7 The general steps included in the data sharing processes for EOL and CyberSEES.

For EOL, the data sharing processes show how data can be shared across research and public community and start with the EOL human mediators' effort to seek diverse data providers and getting an agreement from them to share their data. EOL and the data provider's human mediators then collaboratively set up the formal data sharing partnership and the technology data sharing connector. Last but not least, after the first time successfully sharing data, the human mediators need to maintain a partnership by updating the data regularly. How each step was carried out varies a lot from one content partner to another depending on who the content partner human mediators are, what their relationship is with other content partner human actors, and the data sharing contexts of the content partners.

In the case of CyberSEES, the data sharing processes show how Biocubes data created by non-professionals can be shared like data created by researchers. Similar to EOL, three general steps were identified in this case. The Biocubes project organizers prepared both offline and online environment for encouraging non-professionals to collect and share data. Then, with iNaturalist community members' help, the data quality was improved and reached research-grade. Lastly, the pre-existing partnership built between EOL and iNaturalist allows research-grade Biocubes data to be shared on EOL, in the same way to researcher-created data.

Unlike EOL, where the three steps were closely connected and carried out in sequence, the steps identified in CyberSEES were loosely connected and not carried out in a certain order. Before the Biocubes data were collected and shared on iNaturalist, the formal

partnerships between EOL and iNaturalist had already been long established. In addition, when iNaturalist community members could review Biocubes data was unpredictable; in other words, improving data quality is a long tail process. Even though the steps are not sequential for CyberSEES, it is essential that each of them be completed so that Biocubes data are shared on EOL.

### 6.3 Implications for theory: connecting data sharing contexts and processes

In Chapter 2, this dissertation adapted the theoretical frameworks developed in previous studies into two theoretical frameworks (Figure 6.3): the framework of data sharing processes, adapted from Fecher et al.'s (2015) framework of data sharing, and the framework of data sharing contexts, adapted from Thornton et al.'s (2012) model of microfunctions of institutional logics. In addition, another two models built in previous studies were adopted to help better understand the core concepts of the two adapted theoretical frameworks: the data life cycle model (Rüegg et al., 2014) added clarity to the “data” in the framework of data sharing processes and the organizational identity model (Ashforth et al., 2008) helped to understand the “organizational identities” of the data providers and the human mediators in the framework of data sharing contexts.

These two adapted theoretical frameworks provided theoretical guidance on sampling the cases and conducting initial qualitative analyses to investigate the data sharing practices across research and public communities. The analyses' results, in turn, provide a deeper understanding of the frameworks. Based on this deeper understanding, the two frameworks were integrated into one framework.

The goal of developing this new integrated framework of data sharing is to provide an overview of an ecosystem of a knowledge infrastructure that supports people sharing data widely across research and public communities in online environments. Part of this ecosystem, especially the social interactions between human mediators, was invisible in previous studies on data sharing.

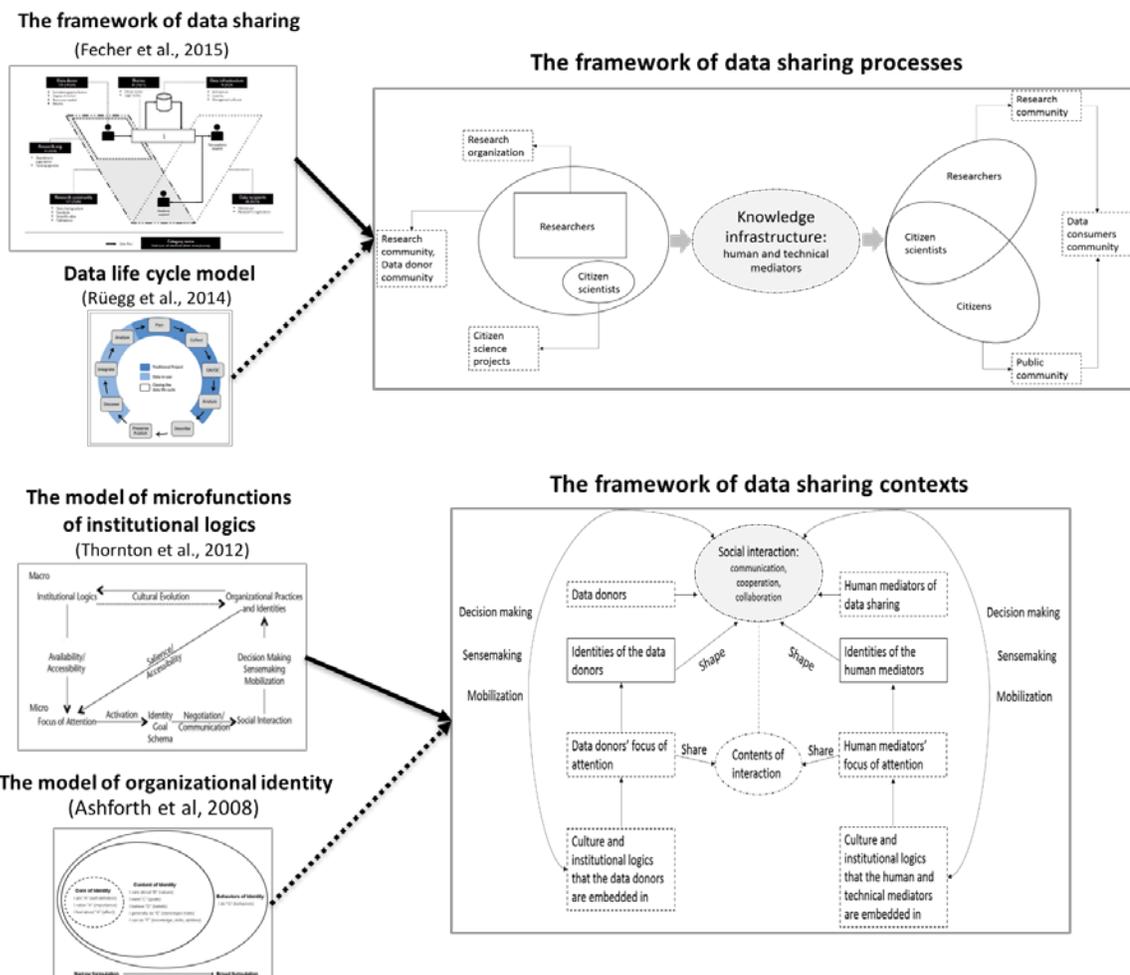


Figure 6.3 Adapting the framework of data sharing, model of data life cycle model, the model of microfunctions of institutional logics, and the model of identification into two theoretical frameworks: the framework of data sharing processes and the framework of data sharing contexts.

### 6.3.1 First version integrated framework of data sharing

Figure 6.4 shows the first version of the integrated framework of data sharing. It was developed based on the preliminary investigation of the two cases. As discussed in Chapter 2, the framework of academic data sharing (Fecher et al., 2015) focuses on the research community and individual-level data sharing where individual researchers are influenced by internal factors from themselves and external factors from the research organizations and communities. In this first version of the integrated framework, the public community was incorporated into the research community and a distinction was made between individual-level and collective-level data providers. Additionally, the specific existing forms of the two different levels of data providers are clarified. The details of this framework will be explain as follows.

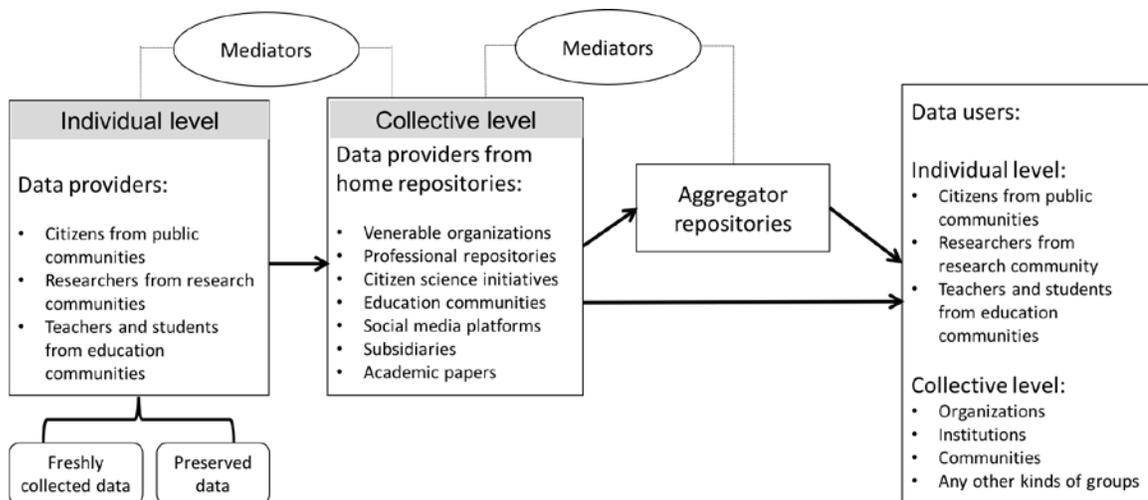


Figure 6.4 The first version integrated framework of data sharing processes and contexts.

#### 6.3.1.1 Individual-level data providers

This dissertation identified both individual-level and collective-level data providers who generate and share biodiversity research data in online environments. The individual-level data providers could belong to research community and/or public community.

When the individual-level data providers belong to the research community, they share the core value that data are generated and shared for solving scientific problems. In this circumstance, on the one hand researchers are responsible for their own data collection and sharing behavior. On the other hand, their data collection and sharing behaviors are important constituent parts of organizational data sharing practices in that “researchers” and “scientists” are organizational identities rooted in research institutions, government, non-government organizations, or other types of formal organizations in research community.

When the individual-level data providers belong to the public community, there could be two conditions. The first is that the public community does not have any specific and explicit shared interests in a certain knowledge domain or non-research profession and there is a large variety of scenarios in which the individual-level data provider could collect data and share data. The micro-level examples of this kind of public community could be Wikipedia, YouTube, or Vimeo.

The second condition is that the public community has explicit shared interests in a certain knowledge domain or non-research profession. The meso-level examples of this kind of public community could be a community of performing arts or an education

community. The individual-level data providers in this kind of community focus on achieving different goals from contributing data to solving research problems. If they have an opportunity to collect and share biodiversity data, it would be unlikely that creating and sharing data become their primary focus. For example, Biocubes project participants were from the education community and their identities as educators were activated when collecting and sharing data. In this case, it made sense to learn how to implement the Biocubes project in their teaching, rather than focusing on exploring how to share good quality data.

However, a person can have multiple identities in different social situations. For example, a researcher can become an individual-level data provider from the public community when s/he takes a beautiful bird photo on his/her vacation and shares this photo on Flickr. This photo is then aggregated by EOL and is discovered and used by an ornithologist in his/her research. Therefore, what the individual data providers' identities are at the moment they collect and share data is the criterion to categorize them into different communities. These identities shape their data collection and sharing behavior.

There are differences in the awareness of individual data providers generating research data from research community as opposed to the public. For individual-level data providers in the research community, when they are collecting data they are clearly aware and have no doubt that they are research data and are collected for use in research. However, for data providers from the public community, they might not be aware that their actions can be considered generating research data or the data can be used as

research data; they may think they are doing something else not related to research at all. For example, posting a photo or video on a social media platform or editing a Wikipedia article are originally non-research data, but may become research data later.

### 6.3.1.2 The stages of the data life cycle

In the framework of academic data sharing (Fecher et al., 2015), it is not clear at what stage of the life cycle data are shared with others. The two case studies found that data providers could share both freshly collected and preserved data. Freshly collected data are new data that have been generated recently or is real-time data and might not have been fully analyzed or published in academic papers or other formal reports/publications. The amount of freshly collected data could keep growing every day.

Preserved data is that which has been generated and stored for a while or even a long time. Collection and uploading to public online environments does not usually happen close together in time and the data wait a longer time than freshly collected data for it to be ready to be uploaded. These data might have been shared locally and/or privately with other researchers on a small scale and, if initially collected for scientific purposes, might have gone through analysis and description processes, and very likely the publication process as well. However, if these data were initially collected for non-research purposes by non-professionals, they might remain in the local storage of the data creators' digital devices or personal computers. The data creators did not yet find a good time and/or reason to upload them to an online environment.

### 6.3.1.3 Collective-level data providers: the home repositories of the data

Irrespective of why individual data providers collect data and within what community, if they decide they are going to share their data publicly online, they need to make the decision upon what online environments (i.e., websites, platforms) they want their data to be seen. These online environments become their data's home repositories as reported in the findings in in Chapter 4 and 5. It is not uncommon that individual data providers share their data on multiple independent platforms. Therefore, data could appear in multiple home repositories and any further sharing of their data would start from these home repositories.

In the case of EOL, seven different types of home repositories (i.e., content partners) have been identified: venerable organizations, professional repositories, citizen science initiatives, social media platforms, subsidiaries, and academic papers. Most of these home repositories contain data that are contributed by more than one individual-level data provider from the same or different project(s) or organization(s). A small number of home repositories were built for sharing only one individual-level provider's data, but are still described as a database, a project or even a community, rather than using the personal identity of the individual provider (e.g., the data providers' name). In this way, the identities of these home repositories are usually not much different from home repositories that contain multiple data providers' data.

The identities of most home repositories are dehumanized or depersonalized no matter how many individual-level data providers contribute data to these repositories. The

identities of the home repository often refer to the products (i.e., database, project), rather than the human actors themselves. This depersonalized identity is more appropriate for building formal partnerships because it helps improve the social trust between the human collaborators (Brewer, 2008). Therefore, all these home repositories are considered collective-level data providers. The data contributed by individual-level data providers go through a depersonalization process by being uploaded to the home repository and being shared further to a wider range of online environments.

#### 6.3.1.4 Mediators

In Chapter 4 and 5, how identities changed from data creators to data providers was clarified: only if data creators are willing to share their data and their data are actually shared can they be considered data providers. If the data are kept by the creators themselves, they are not yet data providers. These data creators need to have help from either human and/or technology mediators to share their data with others in the online environments; data sharing does not happen on its own.

In the two cases described in this dissertation, for individual-level data providers, the mediators are responsible for making their data available to users first on the home repositories and then on the aggregator repositories. For collective-level data providers, the mediators are responsible for transferring the data to the aggregator repositories and making it available to users.

All research data are generated under certain circumstances and biodiversity data is no exception. Given that researchers in the research community create the circumstances in which they generate data, the mediators' responsibilities mainly focus on encouraging them to share data, providing online environments and tools for them to share data, and transferring their data from one platform to another. However, because individual-level data providers in other communities need other people (e.g., human mediators) to create the circumstances for them to generate data or for turning the data into research data, these tasks should also be included in the mediators' responsibilities.

#### 6.3.1.5 Data users

Although this dissertation did not directly address issues related to data users, when investigating who the data providers are, who the data mediators are, and what the data sharing processes are, an understanding about who the data users are was gained as a valuable side product. In the same way that there are different levels of data providers, there are also different levels of data users: individual-level data users and collective-level data users.

#### 6.3.2 Second version of the integrated framework of data sharing

The first version of the integrated framework is a nascent illustration of the contexts and the processes of data sharing. It includes the high level details of different level data providers that were found in EOL and Biocubes. The second version of the framework was developed and is more generalizable. It clearly reflects the interrelationships between different levels of data provider (i.e., individual-level and collective-level data

providers), different levels and types of data sharing mediators (i.e., human mediators, technology mediators, first layer mediators, second layer mediators, third layer mediators), and the organizations, institutions, and communities they are embedded in. The second version of the framework presented in Figure 6.5 provides an overview of the relational and ecological system of the knowledge infrastructure that supports data sharing across communities in online environments. Each part of the framework will be introduced in the remainder of this section.

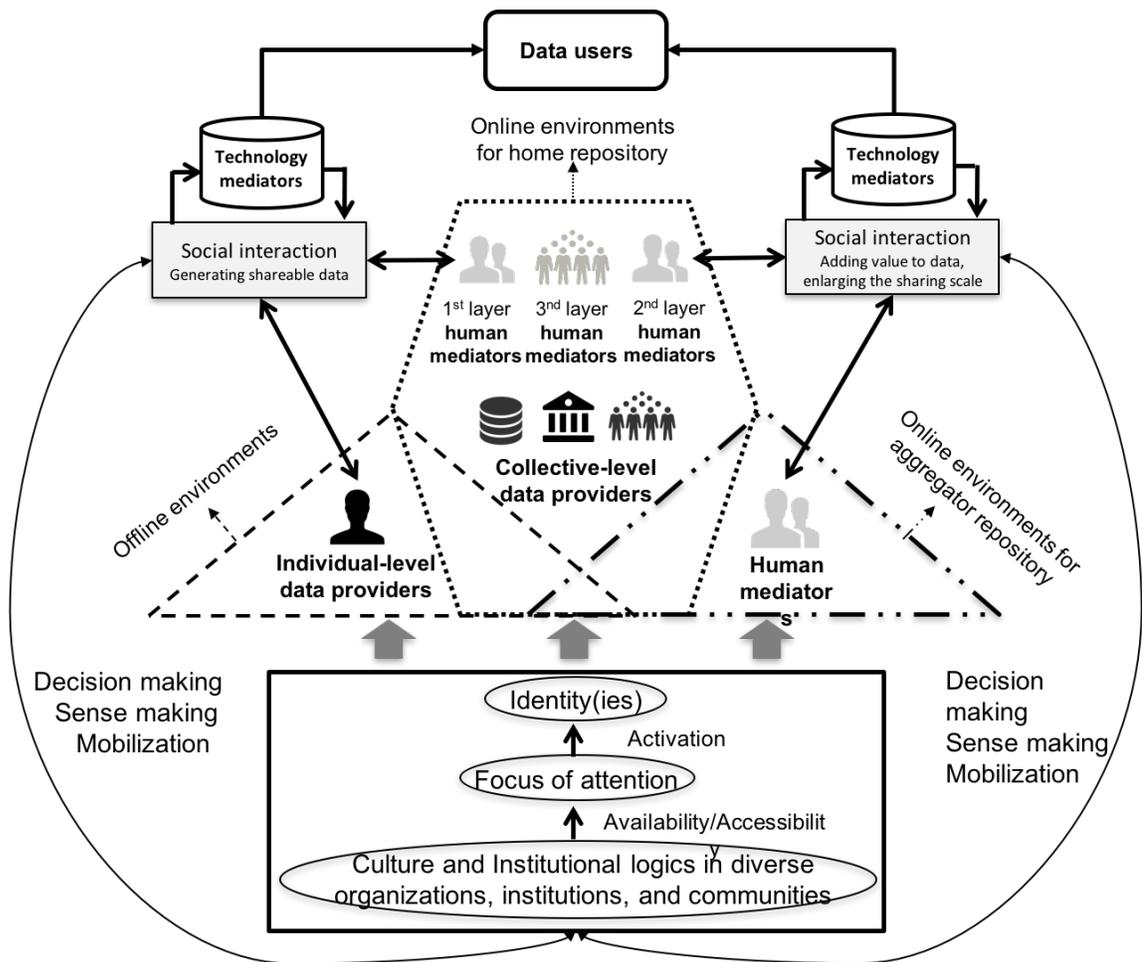


Figure 6.5 The second version integrated framework of research data sharing.

Figure 6.6 shows the three types of data sharing environments included in the framework: offline environments, online environments for the home repository, and online environments for the aggregator repository. The environments refer to any circumstances and social situations in which the human actors play meaningful roles in sharing data. In these data sharing environments, there are individual-level data providers—individual human actors who first exist in various offline environments—and collective-level data providers—databases, organizations, institutions, communities, and any other groups that exist in both offline and online environments. When they are in the online environment, they represent different existing forms of home repositories.

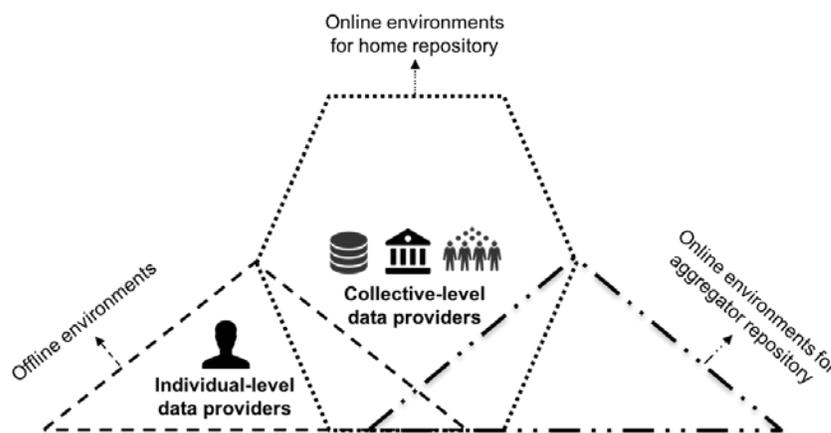


Figure 6.6 Data sharing environments and different levels of data providers.

Below the three environments is the “soil” (Figure 6.7). The soil provides the environments, the human actors, and technology a medium in which to grow and share data. Its key nutrients are the culture and institutional logics of the organizations, institutions, and communities which shape the relationship between the human actors, technology, organizations, institutions, and communities. The human actors’ focus of

attention and organizational identities when sharing data are decided by what culture and institutional logics are available and accessible to them.

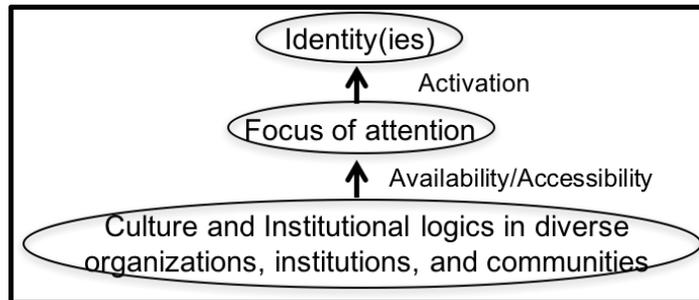


Figure 6.7 Data sharing environments and different level data providers.

A human actor can have multiple social identities. Under different data sharing circumstances and social situations, these identities help people as social actors recognize the single or multiple roles they choose to play and the work they do (Thornton et al., 2012). The three data sharing environments have intersection areas, indicating that human actors could have multiple co-existing roles and, if so, are functional in two or three of the three environments.

Figure 6.6 and Figure 6.7 introduce the data sharing contexts. Figure 6.8 shows the data sharing processes. The icons of human mediators and the backgrounds of social interaction are colored grey because they are usually invisible in the ecosystem of data sharing knowledge infrastructure.

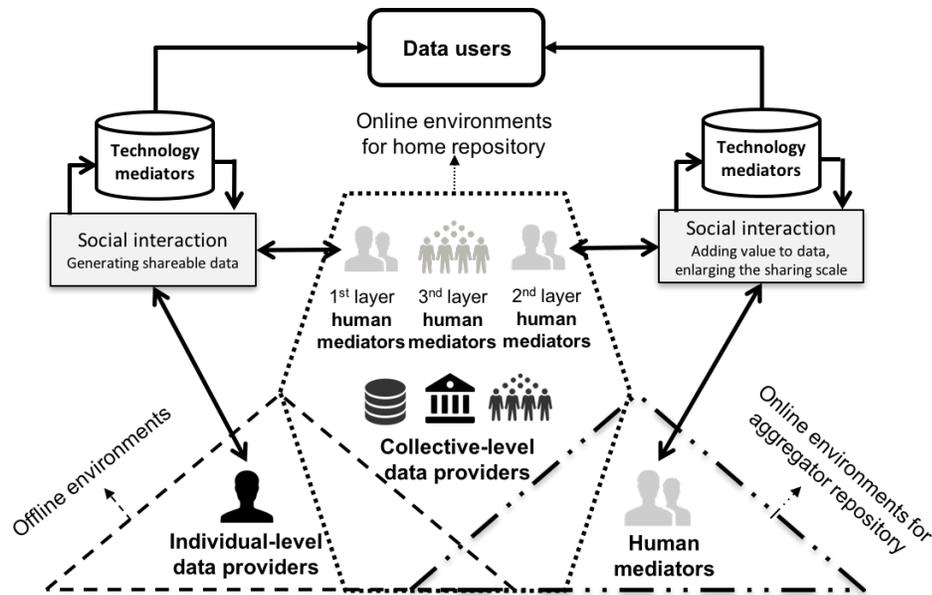


Figure 6.8 Data sharing processes facilitated by data sharing mediators.

Human and technology data sharing mediators facilitate data sharing done by individual-level and collective-level data providers. First layer human mediators prepare offline and online data sharing environments for individual-level data providers to share data with technology mediator support. In addition, these human mediators are responsible for communicating, collaborating, and cooperating with individual-level data providers to facilitate their data sharing so that their data can be available on home repositories via technology mediators for the first time. The social interaction between data providers and human mediators could happen in online and/or offline environments.

After individual-level data providers' data are shared successfully in the home repository, second and third layer human mediators facilitate data sharing from the home repository to the aggregator repository. These mediators are responsible for communicating, collaborating, and cooperating with data aggregator repository human mediators to build

data sharing partnerships and the technology data connector (part of the technology mediators) between the home repository and the aggregator repository. Second layer human mediators could represent the home repositories, or in other words, the collective-level data providers.

Third layer human mediators are responsible for ensuring data quality in the home repository. When data creators are non-professionals, these mediators could be a community of non-professionals with different data expertise. When the data creators are researchers, the third layer human mediators are still needed for ensuring the data quality because the data aggregator repository will not accept a data source without reviewing its reliability and trustworthiness. Reviewing data created by researchers requires that the third layer of human mediators be experts. After the data connector is successfully established and the data reach a certain level of quality (i.e., research-grade), that data can be shared from the home repository to the aggregator repository. New value is added to the data in the aggregator repository because they are shared in a wider range of online environments. Any data users from both the research and public community can access the data via technology mediators.

No matter whether the human actors are data providers or mediators, their identities shaped the social interactions (i.e., communication, collaboration, cooperation) via decision making, sensemaking, and mobilization. These social interactions in turn influence culture and institutional logics as time went by. This relationship is highlighted in Figure 6.9. Both interviewees from the first case said that compared to earlier years,

data sharing had become easier and more acceptable among both the existing and new data providers as time went by.

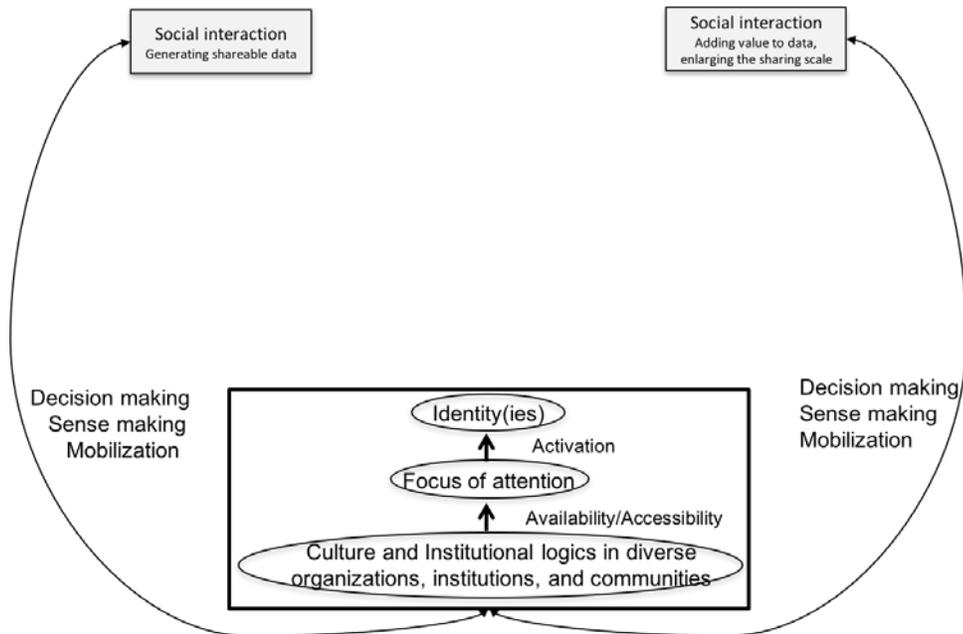


Figure 6.9 The relationships between identities and social interactions.

The in-depth understanding of EOL and CyberSEES enables the development of a new integrated framework of data sharing. It also provides valuable practical implications for data sharing. The following sections will first introduce implications for the data sharing practices, and then implications for the design of data sharing.

#### 6.4 Implications for the data sharing practices

From data being created in an offline environment to being shared in a home repository and then aggregated in an aggregator repository, this dissertation identified a series of efforts made by human actors for sharing research data across research and public

communities. Among all these efforts, the collaborative efforts made by human mediators are the key that enable data to become available for both researchers and non-professionals. Any missing links among these collaborative efforts would block the pathway of sharing data effectively with both research and public communities.

These collaborative efforts are critical data sharing practices, indicating that the data sharing process across research and public communities is a complex socio-technical process. This socio-technical process is composed of a series of activities driven by the human mediators. The activities happen at organizational and/or personal levels and are influenced by the human mediators' identities and technology resources that are available for them. The activities revealed in EOL and CyberSEES not only include most types of human work furthering the aggregation and sharing of data reported in previous research on developing cyberinfrastructure (e.g., Lee et al., 2006; Bietz et al. 2010), but also reveal new types of human work.

When put in an infrastructural context, previous research indicates that sharing data within a research community via data repositories for the purpose of advancing research work is an ultimate goal and essential part of developing a cyberinfrastructure (Lee et al., 2006; Bietz et al., 2010). It is widely accepted that the concept of cyberinfrastructure restricts the development of it for the research community and research purposes, but not for the general public for non-research purposes. Similarly, when not being put in an infrastructural context, previous studies on research data sharing and data repositories mainly focus on data sharing by and for researchers, not non-professionals. It has been

rooted deeply in people's mind that research data are typically created and used only by researchers. However, since the speedy development of citizen science, there are increasing numbers of non-professional citizens become research data creators (e.g., of citizen science data) and users.

Before this dissertation, it was unclear how research data can be shared and aggregated not only for the research community, but also for the public community; more specifically, how research data created by non-professionals can be shared and aggregated like those created by researchers. In this dissertation, the data sharing practices illustrated by the example cases do not only target the research community, but also the public community. Besides promoting research, the core missions of both EOL and CyberSEES include sharing research data, information, and knowledge with the public community.

The difference between targeted data creators and user communities among the cyberinfrastructures and data repositories previously studied and those studied in this dissertation are that whether these communities belong only to the research community or to both the research and public communities. Nonetheless, the human workers in cyberinfrastructure, data repositories, and EOL and iNaturalist share the same goal: aggregating research data from different data providers and provide public access to it to potential data users. This indicates that sharing research data publicly, whether for the research and/or public community, requires human workers to make similar and certain types of efforts.

However, there are differences in the human efforts between sharing data with the research community only and with both research and public communities. Compared with cyberinfrastructures and research data repositories, the human mediators in EOL and iNaturalist need to work with more diverse data providers in terms of their data sources and management style (e.g., data providers who provide data created by researchers, by non-professionals, or by both). These human workers also need to display the data and provide access to it in a more readable and understandable way, so that non-professionals can obtain and understand easily it.

#### 6.4.1 Data sharing practice for research community and for both research and public communities

In general, the socio-technical process of data sharing revealed in this dissertation can be considered a synergizing process, as introduced in Chapter 2 in relation to developing a cyberinfrastructure (Bietz et al., 2010). In both example cases, the human mediators were involved in the two sub-processes of synergizing (Bietz et al., 2010): leveraging and aligning various social-technical relationships at both collective and individual levels. Relationships among human actors, organizations, and technologies could be created and maintained for the purposes of first creating data in an offline environment and then sharing it successfully across research and public communities in multiple online environments.

To be specific about the exact activities of data sharing, this dissertation divided them into three steps in each case. The process for existing data being aggregated in EOL

included: 1) preparing social relationships and reaching a mutual agreement between EOL and its content partners; 2) developing a data sharing connector to transfer data from content partners to EOL; and 3) updating data or the data sharing connector to keep the data fresh and increase its quality. The process for data being created and shared in the online environments for CyberSEES included 1) the origin of data in an offline environment; 2) the origin of data in the first online environment (i.e., home repository); 3) sharing the data to an aggregator repository. These activities were accomplished collaboratively by human workers as data providers and data mediators. Compared with the previous cyberinfrastructure studies, there are similarities and differences among specific data sharing activities.

For example, in Bietz et al.'s (2010) study, building a community repository for data sharing is a key activity of developing the studied cyberinfrastructure. They reported three perspectives to the activity of data sharing: importing data, metadata, and landscaping data (Bietz et al. 2010). When building a repository for this cyberinfrastructure, this study reported how one programmer worked on creating the data schema and getting access to a small number of datasets as test cases. A senior administrator helped introduce the scientists (i.e., data creators) behind these datasets to the programmer. These scientists' grants required them to publish their data through this repository. The programmer then parsed and imported their data into the repository and resolved technical issues of inconsistent data format and standard. But the programmer himself could not solve the issue of creating a metadata schema because of a lack of agreement about metadata among the scientists. Instead, this cyberinfrastructure project

had to join an organization that creates and provides data standards and agreements among scientists in a certain research area. The project members then decided to share the same data in multiple systems, linking the local data to the global data by creating alignments with other systems. Ideally a scientist could share their data with the current repository first, and then decide whether they want to click a button to submit their data to another system.

The data importing activities in Bietz et al.'s (2010) study are most similar to developing a data sharing connector between content partners and EOL. The social relationships between the programmer and the scientists who own the datasets in Bietz et al.'s (2010) study are similar to those between one type of EOL content partner, subsidiaries, and EOL. These subsidiaries type content partners belonged to EOL fellows who received grants from EOL, were managed by the SPG coordinators, and was required to follow EOL's data policies. However, the technology relationships between the programmer and the scientists who own the datasets in Bietz et al.'s (2010) study were different to those between EOL fellows and EOL. In Bietz et al.'s (2010) study, the scientists provided inconsistently formatted datasets directly to the programmer. The EOL fellows are required to adopt an existing system (i.e., Lifedesk) to share data locally first and the export a standard data file that can be easily imported into EOL. That said, it is common for EOL technicians to encounter the same situation with other types of content partners when working on their databases and data files; in this case, they need to do extra work to transfer the data format to an EOL compatible one.

Besides the social relationships between EOL subsidiaries type content partners and EOL, other types of social relationships were also revealed between other types of content partners and EOL. These other types of content partners (e.g., venerable organizations, professional repositories, citizen science initiatives, social media platforms, education communities) are independent from EOL, meaning they are not managed or owned by EOL or required by grants they receive to share data with EOL. The human mediators, the SPGers, and the EOL technicians engaged in significantly more complex and significant relationships leverage and alignment work to connect diverse types of content partner; the different types of content partner have completely different histories, needs, available sources, and visions. Therefore, there is a significant difference between data sharing practices in a cyberinfrastructure study that involves the research community only and EOL that targets both research and public communities.

This complex relationships leverage and alignment work is also reflected by the different possibilities for the identities of the data providers' human mediators (Table 4.1).

Previous studies on cyberinfrastructure (e.g., Lee et al., 2006; Bietz et al., 2010) rarely identify the specific identities of the data providers or simply identify them as data contributors (i.e., researcher, scientists). However, this dissertation shows that in the case of EOL, the data sharing process varied from one content partner to another because of the influence of the identities available to the human mediators from the content partners.

Who the available data providers' human mediators that work with EOL human mediators are influences the how the social-technical relationships are created and maintained. The influences of these identities will be discussed in more detail below.

The technology relationships between EOL's different types of content partners and EOL are influenced by these content partners' technology intelligence and other available resources. Although EOL was improving its data file schema and adopted a widely accepted biology data standard (i.e., DwC-A), it still supports different approaches of developing data connectors based on different data providers' needs.

In addition, it is uncommon to see the human efforts behind collecting data and initially sharing it in an online environment reported in previous cyberinfrastructure studies. It has been taken for granted that the researchers should follow certain scientific methods to collect and analyze data and ensure its quality. Human efforts are focused more on aggregating and sharing existing scientific data. Therefore, the second case in this dissertation, CyberSEES, revealed the data sharing practices for freshly collected citizen science data being shared from its home repository to an aggregator repository. The data sharing practices in CyberSEES show the process of linking local data to global data from the perspective of sharing across research and public communities.

During the data sharing process, tensions between local citizen science level and global research level data sharing were resolved by human mediators at different levels (i.e., collective and individual levels). This indicates that an infrastructure occurs as local citizen science data collection and sharing practices can be afforded by large-scale technology (Star & Ruhleder, 1996). This also inspires future cyberinfrastructure development to consider how to better include data that created not only by researchers,

but also by non-professionals, as well as sharing data for not only the research community, but also the public community. How the tensions between local citizen science level and global research level data sharing were resolved are discussed below.

#### 6.4.2 Sharing data on the home repository and on the aggregator repository

The previous section discussed similarities and differences between data sharing practices within the research community and across research and public communities in the context of infrastructure. This section will focus on the data sharing practices of the home and the aggregator repository from the perspective of the difficulties of sharing data, which is meaningful for facilitating sharing data both within research community and across research and public communities.

Previous research has found that the largest discrepancy in data sharing practices is between researchers' willingness to share and whether sharing actually happens (Carlson & Stowell-Bracke, 2013; see also Tenopir et al., 2015). The major reasons for not sharing is that documenting and sharing data always takes time and effort that are not appropriately rewarded (Kratz et al., 2015) and that there are perceived risks of sharing (Tenopir et al., 2015). In the case of EOL, content partners have already shared their data with others, with the sharing not just limited to private sharing since most content partners allowed their data to be shared in online environments (i.e., home repositories) that are publicly viewable. In other words, these content partners have overcome the difficulties of sharing data and successfully shared data before sharing it with EOL.

Nevertheless, it still took a long time and a significant amount of human effort to share these data publicly on EOL.

Even though the data have been shared publicly in these online environments, it is still very hard to aggregate, share, and/or publish these data on another platform. The question is why. A basic answer may be similar to the reasons people choose not to share regardless of whether they are willing to share: sharing data in another platform takes time and effort that is likely not rewarded appropriately together with perceiving risk associated with sharing. This is not an incorrect answer. The EOL findings demonstrated that building data sharing partnerships was time consuming and required a large amount of human effort. However, on deeper inspection, the efforts made sharing data on an aggregator repository like EOL are different from the efforts made sharing data on home repositories like iNaturalist. The differences are reflected first by *whose time* and *what efforts* data sharing actually takes; and second by at *what level* data sharing happens? These differences could be associated with the difficulties of aggregating, sharing, and/or publishing the already publicly shared data on a different platform.

Previous studies on understanding data sharing practices have focused on micro-level sharing (i.e., individual level), referring to individual researchers' sharing motivations and behaviors (e.g., Fecher et al., 2014; Tenopir et al., 2015; Kratz et al., 2015).

Common mechanisms or channels the individual researchers used to share data were email/direct contact, personal website, journal website, and repository (Kim & Stanton, 2012; Kratz et al., 2015). Sharing data on personal websites and journal websites are the

practices of sharing data on home repositories. Sharing data via email/direct contact and on a repository can be either the practices of sharing data on a home repository or sharing data with an aggregator repository.

However, it is not clear what roles or organizational identities these individual researchers actually played or adopted when sharing data regardless of whether it was on a home or an aggregator repository, or whether their sharing behavior represents just themselves or a group (e.g., research teams, organizations, or communities). In addition, it is not clear what efforts or what processes these individual researchers actually made for successfully sharing data, for example whether they undertake data sharing work by themselves or get help from colleagues, graduate students, technicians, or data managers?

Schmidt et al., (2016) distinguished different data professional roles in their survey conducted to understand people's perceptions of the term "open data." These roles included data users, researchers, data scientists, data managers, and technologists. Among the 1248 participants in the survey,

"82.3% (1025 respondents) saw themselves as data users, 57.6% as data providers (718 respondents), and 25.3% as data managers (315 respondents) (multiple answers were allowed). About 5.3% (66 respondents) of all respondents saw themselves in other or multiple data roles, and/or were unsure on how to classify themselves, e.g. as researchers, (data) librarian, software developer, administrator etc." (Schmidt et al., 2016, p. 3).

These results indicate that individual researchers' data sharing practices might include hidden efforts made by different data professional roles. The persons who assume these professional roles might or might not be the same persons.

That participants claim multiple roles in Schmidt et al., (2016) is supported by the findings in this dissertation. Some human mediators of data sharing from the EOL content partners adopted more than one identity when working with EOL to build a data sharing partnership, whereas others adopted only one role at a time. These roles are reflected by social identities: what identities are adopted by the human mediators are decided by the organizations, institutions, and communities they belong to. Identities are activated by the specific social situations for data sharing.

The social situations for sharing data in a home repository versus an aggregator repository are different. This dissertation found that sharing data in a home repository is usually carried out by individual-level data providers at a relatively small scale; sharing data in the aggregator repositories is likely carried out by collective-level data providers at relatively large scale. Sharing data in a home repository usually correlates with preparing and uploading the data themselves, or sometimes with a technician's help.

However, when it comes to sharing data in an aggregator repository, the data provider may or may not be the same data provider who uploaded data directly to the home repository. For example, data providers are not individual data contributors, but instead data managers who are responsible for managing a repository containing a group of

individual data providers' data. These data providers shared their data publicly in the online environment provided by the home repository once already, so need to evaluate the request to make their data available on a different platform. They might consider whether they want to share their data together with many other data from different sources they might not be familiar with. Whether they need to make extra effort to prepare their data in a different way from the home repository may be another consideration. Regardless of whether these efforts are rewarded, there are additional internal and external factors that could influence a data provider's decision to share or not in an aggregator repository, as compared with sharing data in a home repository.

For an aggregator repository, requesting data from the home repositories is likely to involve more human mediators to make data sharing decisions compared with requesting data from individual-level data providers. Undoubtedly, there is a higher communication cost when there are more human actors involved in decision making, especially when the home repository represents any large organization, institution, or community with relatively complex organizational and power structures. In addition, there might also be more technological demands to make the large scale data transfer from one platform to another.

The identities held by the human mediators from data providers influence data sharing practices significantly. Like getting working with a home repository, it could also be more difficult and complex to get approval from and collaboration with an organization, institution, or community to aggregate their data than from an individual-level data

provider. As the data quantity increases, there might be greater requirements for the technology as well: there have been previous studies focused on investigating technological difficulties of aggregating data. However, this dissertation focuses on the difficulties of human collaboration to share data in the social situations of sharing data with aggregator repositories. Difficulties discussed above are inevitable when considering sharing data in a wider range of online environments, which have to be overcome by an experienced human mediator with enough domain knowledge, technology background, communication skills, and patience.

#### 6.4.3 Peer reviewing the data created by non-professionals

Given the tensions between citizen science data sharing and global research level data sharing, how tensions over data quality was resolved by aligning human mediators' and technology's power will be discussed here, so that the citizen science data can be shared in a global research data aggregator repository together with research data created by researchers.

The second case in this dissertation, CyberSEES, focused on a citizen science project in which the research data were collected by a majority of educators, in other words non-professionals in the public community. This is true for EOL as well, because content partners provide data that are collected not only by researchers, but also non-professionals who are not specialists in a biology related discipline. A similar phenomenon has also been noted in other large-scale data hubs in which citizen science data sources are included in order to have more comprehensive and less biased

aggregated data (Otegui et al., 2013; see also Daume & Galaz, 2016).

Many previous studies on data sharing practices have focused on how human actor, mostly researchers from different subject disciplines, came from academia to share data (see Schmidt et al., 2016). Few studies have focused on how human actors not from academia and not researchers share data, despite plenty of studies that have focused on encouraging non-professionals to participate in citizen science projects and developing various platforms and applications to support them in collecting and sharing data.

Researchers and scientists running these studies usually request data collected by non-professionals to be shared on a home repository. However, relatively little focus of these studies has been put on the practices related to how research data created and shared by non-professionals and that created and shared by researchers are shared together in a larger online environment.

The results in this dissertation provide a valuable example of this kind of data sharing practice and illustrate how research data contributed by both researchers and non-professionals can be considered equally trustful and reliable when shared. In Biocubes, the “magic” practice that enables the citizen science data to be shared with professionally collected data on EOL is helping the data reach research-grade on the home repository, iNaturalist. This help comes from two types of data sharing human mediator: firstly, the scientists who were involved in the citizen science project, Biocubes, as organizers and iNaturalist community members and, secondly, other iNaturalist community members not related to the Biocubes project. Besides the fact that they are all iNaturalist community

members, the other most important common characteristic of Biocubes scientists and iNaturalist community members who helped improve the data quality is that they all played the role of data reviewers. They reviewed the data and provided its results (i.e., agreeing on ID, suggesting a new ID) on the Biocubes data pages on iNaturalist. These review results are publicly viewable and can be considered evidence of the review processes. It is only by having these human mediator's review results that the data have the possibility of becoming research-grade. It is important to ask why having community members help review the data is important and considered a required step.

In the research community from the data users' perspective, researchers expect data published online are validated in ways that they trust (Kratz et al., 2015). The peer review method is most trusted and valued by researchers compared with other methods, such as having a traditional paper as the basis of a dataset, the data having been successfully reused by others, and the data having been described in a data paper (Kratz et al., 2015). Although having iNaturalist community members review the data is not strictly a peer review process, it shares the spirit of peer review. In the research community, peer review is most usually a group of experts in the same subject discipline who evaluate a scholarly work and determine whether it is appropriate to be published. In the iNaturalist community, a scholarly work is replaced by one observation data, the group of experts is replaced by the community members, and the determination of "accept" is replaced by "research-grade."

It is worth mentioning two points. First, to ensure the data quality what is reviewed is

one observation datum, not “a database.” In Kratz et al.’s (2015) paper, having a traditional paper as the basis of the whole dataset is not as trusted as direct peer-review of the data themselves. This could explain why some citizen science projects publish their data collected by the non-professionals in peer-reviewed journal. However, their datasets might not be accepted by the research community at large and cannot be shared in the same venue as data collected by researchers. Reviewing each datum by a human actor is a trustworthy method to increase reliability of the data and facilitating further sharing of it. For data that are rooted in a public community, conducting a peer-review style evaluation can be a key data sharing practice for making them trustworthy enough to be further shared together with traditional research data in a wider range of online environments. The reviewers of the data created by non-professionals can be considered a group of data sharing human mediators with loose ties.

Second, the determination of “research-grade” on iNaturalist is not only made by iNaturalist community members, but also by an algorithm. However, without the iNaturalist community members’ input, the algorithm could not “accept” the observation data as research-grade level data. Although there is the potential that some of the human review could be replaced by technology-mediated review, as demonstrated in Sullivan et al. (2014), the human effort in the peer-review process still could not be replaced because the machine is not yet smart enough to replace the complex cognitive functions of the human brain needed to validate research work/data.

In addition, on EOL, data created by non-professionals are not only from citizen science initiatives, but also social media. Non-professionals on social media are not intentionally contributing research data like citizen scientists do, but instead share content that has potential to become research data, even if they are not necessarily aware of this at the time. Daume & Galaz (2016) produced some of the first research on this phenomenon on Twitter. They noticed that some Twitter users upload real-time information about biodiversity observations with species determination requests and, with other Twitter users help, 86% of the total 191 Twitter message samples received at least one suggested identification from another Twitter user, 76% of which are correct. These Twitter samples, especially ones with correct determinations, are considered valuable ecological monitoring data (i.e., research data).

The difference between iNaturalist and Twitter is that the former is a community built specifically for naturalists to upload only biodiversity observation data (i.e., occurrence data), whereas but Twitter is a community built for anyone who has Internet access. iNaturalist highlights that these data could be used as research data by scientists and promotes this belief to the entire iNaturalist community. Twitter users who upload biodiversity observation data are not collecting the data for research purposes—these users usually share many other kinds of information about topics other than biodiversity observation data. However, regardless of this difference, the process of getting data verified is similar to that on iNaturalist: community members help determine, through the Twitter platform, the identification of each datum. This help from the community is extremely valuable and has huge potential to be further developed and studied in the

future because there will never be enough biodiversity researchers and experts to help review the amount of biodiversity data created by non-professionals. The data created by “the crowd” have to rely on the crowd itself to finish the “peer-review” process that can be only done by people.

## 6.5 Implications for the design of data sharing

In both cases in this dissertation, technology mediators support human mediators effectively transferring data from providers to users. Based on the understanding of technology mediators in both cases, there are a few design suggestions that could help improve technology mediators for facilitating data sharing.

In both EOL’s and iNaturalist’s infrastructures, three sub-systems as key components of the technology mediators were identified: the data provider management/data entry system, the data exhibition system, and the data export system. The data provider management system connects the data providers with the entire infrastructure. The data exhibition system and data export system connect data users with the entire infrastructure. Data providers and users are connected via the technology mediators. The data provider management/data entry system allows data providers to share data by uploading the data files to EOL/uploading the data to iNaturalist. The data exhibition system publicly displays the elements of the data under a certain structure (i.e., species/taxon/observation organism page) to any potential data users. The data export system allows data users to easily create a query and download a batch of data in a CSV file.

Among the three sub-systems, the data sharing practices discussed above are closely related to the first and second sub-system. For EOL, the data provider management system design led to the human mediators' decision on what and how to prepare for sharing data (e.g., prepare data source files). Then the human mediators from both EOL and the content partners preview the data together on the data exhibition system and improve the data source files until the preview results satisfy both sides. The preview processes are private and only available between EOL and the content partners. For Biocubes, the data entry system influenced how the human mediators guided the project participants to collect and share data: since the human mediators collaboratively review and improve the data quality on the data exhibition system, the processes of reviewing and improving data quality are transparent and public viewable. The discussion here will focus on the design related to the first and second sub-systems and that could help make data sharing more efficient in online environments.

The EOL case findings showed that although sometimes there are small technical problems when using the data management system to upload data files and the data exhibition system to review data, these issues did not become true obstacles of sharing data. The amount of time human mediators took to use these two sub-systems is much less than to obtain agreement over preparing and improving data files to build a successful data connector. The design implications for facilitating the aggregation of data is inspired partially by the first two sub-systems on EOL. Improving the design

could make preparing data files easier and therefore make data sharing in a larger scale online environment (i.e., aggregator repository) more efficient.

On the other hand, the Biocubes findings show that the data contributors and the human mediators need to spend a relatively long time using the data entry system and data exhibition system to prepare the data. Only after each datum reached research-grade on the data exhibition system could the datum can be further shared. Therefore, the design implications for facilitating the creation of shareable data is inspired by the two sub-systems on iNaturalist, and is for better designing these two sub-systems themselves on iNaturalist and other home repositories.

The following sections will introduce more details about the implications inspired by the data provider management/data entry system and the data exhibition system on the EOL and iNaturalist platforms.

#### 6.5.1 Preparing ready-to-use data files

The first implication is inspired not only by the two sub-systems on EOL, but also by the two platforms of the home repositories (i.e., Lifedesk and Scratchpads). The design of home repository should take the mobilization of data among different platforms in consideration so that the data can be shared with the aggregator repositories more efficiently. Although Lifedesk and Scratchpads are targeted at researchers to build home repositories for their data, the implication is for all platforms built for sharing data with both the research and public communities.

Lifedesk and Scratchpads take the promotion of data mobilization as one of its missions (Smith et al., 2012). Taking Scratchpads as an example, in order to support data mobilization the platform adopts DwC-A to guide individual data providers how to upload and manage data on Scratchpads (Smith et al., 2012). Because DwC-A is a preferred data file format for sharing biodiversity data in large scale online environments, individual data providers who upload their data to Scratchpads and use it as the home repository can easily export the data files in the DwC-A format. They are also aware that the home repository of their data encourages them to do so, and they can share their data with a wider range of audiences on different platforms.

Assuming the data provided by the individual data providers are reliable and trustworthy, their DwC-A data files can be accepted and immediately uploaded to the EOL content partner management system and other biodiversity aggregator repositories (e.g., Global Biodiversity Information Facility). There is no need to make further effort to prepare the data files by converting the data file into one of the specific formats EOL can use. This design of Scratchpads—and Lifedesk—allows biodiversity data to be ready to share with different platforms as soon as they are uploaded to these two platforms. This could save a large amount of time and energy for human mediators, especially when individual human mediators take multiple identities (e.g., data contributors, data managers, and technicians) or data providers do not have the IT intelligence to create data files by themselves. Human mediators can then focus more on collaborations to improve interpersonal/interorganizational relationships and on improving the data quantity and

quality. Therefore, data providers who use platforms like Lifedesk and Scratchpads might be more likely to agree to share their data on other platforms.

These two platforms are not without issues, however, such as the appropriateness of encouraging data providers to adopt Lifedesk and Scratchpads to share data, the need to improve their usability, and the challenge of using them to share a large amount of data. These issues are beyond the scope of the discussion here and can be addressed in a future study. What is instead emphasized is that the design of Lifedesk and Scratchpad provides a good example of how to incorporate the mobilization of data sharing among different platforms within their platform design.

As the target users of the Lifedesk and Scratchpad are researchers, it might be easier for other home repository platforms also built for researchers to model their mobilization of data between different platforms on Lifedesk and Scratchpad's designs. However, it might be more difficult for home repository platforms that are built for general public to do this since it is not yet clear whether the leaders of these platforms have realized that the content generated by users can be research data. These leaders need to decide whether it is worth putting attention on helping mobilization of research data across different platforms, after developing and designing strategies to identify data that is or has potential to become research data.

### 6.5.2 Design for supporting collecting and uploading data simultaneously

The second implication is inspired by how Biocubes and its human mediators use the data entry system and data exhibition system on iNaturalist. The two systems on the home infrastructure should support and encourage data contributors to easily combine data creation and data upload so that data would be being shared with the home repositories more efficiently, especially for data created by non-professionals.

In the Florida Biocubes project training workshop observed in this dissertation, the project organizers required the data creators combine data collection and data uploading. Therefore, immediately after the data were collected offline, they were uploaded to the iNaturalist platform. Then the data were shared publicly and became ready to be validated and improved by iNaturalist community members. As soon as the data reached research-grade, they could be shared with traditional research data in larger scale online environments.

This data sharing practice supported by the iNaturalist data entry and exhibition system has the advantage of facilitating biodiversity data created by non-professionals to be shared in an online environment. The more non-professionals that share their biodiversity observation records online, there more data to have the potential to become research level data as the base number of the data increase. Biocubes and the iNaturalist platform provide a good example of how to take advantage of this kind of design.

Collecting data using the iNaturalist app on a smartphone is the most convenient way to combine data collection and uploading. However, a smartphone is not an ideal device to get good quality data, especially for taking clear photographs of a moving organism; the data uploaded through the smartphone app are less likely to get other iNaturalist members' validation (Wiggins & He, 2016). However, as the base number of observation data increases, there would be more data being validated so it is still worth encouraging data creators to use a smartphone to upload the real time data they just collected.

Another limitation of a smartphone is that this kind of device is not always available to everybody for a variety of reasons. Biocubes participants from the public community other than those attending the Florida workshop collected data by using paper and pen (i.e., metadata) and traditional cameras (i.e., media). As these data were not uploaded immediately to iNaturalist after the data collection, it was very unlikely that the data would get uploaded to iNaturalist unless the Biocubes project organizers made effort to push that work forward. This was due to other demands on the data creators' time and energy and the lack of focus on creating research data that participants from this community had. Therefore, future research might consider how to design and develop a portable and affordable device to better combine data collection and upload without relying on smartphone device.

### 6.5.3 Visible metadata and data quality assessment information

The third implication is inspired by the iNaturalist platform. Metadata to the data should be captured when uploading data to the online environment for the first time; this information and other data quality related information should be as transparent and visible as possible to data users. Capturing the metadata means that once the data are collected and uploaded, the data are shared with metadata publicly visible (i.e., data observation and creation time, location, media files, etc.). Metadata and well-defined data quality information are the most important attributes to data users deciding whether to use the data or not (Schmidt et al., 2016).

Like many other social media platforms, iNaturalist supports users to generate real-time content. However, iNaturalist is significantly different from most other social media platforms in terms of the platform design because most social media platforms do not provide an interface that is designed for structuring a certain type of data in a specific knowledge domain. As a social network of naturalists and a biodiversity citizen science initiative, iNaturalist uses the design of its data entry system to help structure the data in the format that is most useful for validating species data (Daume & Galaz, 2016).

The key function of structuring the data is to help data creators know what should be recorded when observing an organism, such as the time, date, and location of the observation, as well as a media file of the species if possible. All this information becomes required when determining the data quality of each datum by a human and machine reviewer and are visual and transparent to any data users. Then the data quality

information, including the human reviewers' review results (e.g., agree or not agree the current ID, suggesting a new ID) and machine review results (i.e., data quality assessment panel), are displayed on the same data page. The design of the data exhibition system on the aggregator infrastructure is not able to show detailed metadata for each datum, but it usually provides the link to allow potential users to track back to the original record of each datum on the home infrastructure. The metadata and data quality information on the original record of each datum are visible and completely transparent to the potential data users. This is helpful to them to become more confident when making a decision about whether they want to use this datum.

## 6.6 Conclusion

This chapter first compared the findings of the two cases in terms of the data providers, human and technology mediators, and the data sharing processes. It then introduced the integrated framework of research data sharing developed from on earlier versions of the frameworks and the findings from the two cases.

This chapter discussed the implications for practice and design of data sharing.

- The implications for practice focused on human mediators, emphasizing the differences between data sharing practices in cyberinfrastructure and the two cases in this dissertation; the influences of the complex identities of the human mediators on data sharing practices at collective and individual levels; and a strategy for sharing data created by non-professionals across research and public communities.

- The implications for design focused on the technology mediators, highlighting design ideas for sharing data to home repositories and aggregator repositories more efficiently; and for helping data created by non-professionals become more trustworthy and able to be shared together with the data created by researchers.

## 7 Conclusion

This dissertation asked an overarching question: *how data are shared effectively across research and public communities?* By answering this question, this dissertation achieved two goals: 1) addressing the knowledge gap about research data sharing beyond the research community in large-scale online environments; and 2) developing the corresponding data sharing framework. This dissertation investigated the practices of effectively sharing data across research and public communities in online environments. The practices are reflected by the data sharing contexts and processes. This dissertation therefore broke down the overarching question into three research sub questions:

- Who are the data providers?
- Who are the data sharing mediators?
- What are the data sharing processes?

Through a comparative and multi-layer case study method of using artifacts, documentation, participant observation, and interviews, this dissertation investigated data sharing practices by answering these three questions. The practices were mainly focused on networking people, (digital) artifacts, projects, organizations, institutions, communities, and technologies for sharing data across research and public communities in the ecosystem of a knowledge infrastructure built for data sharing.

Each knowledge infrastructure consists of human and technology infrastructure. The collaborative efforts made by the human mediators and supported by the technology mediators of data sharing within the human and technology infrastructures respectively

are essential for transferring data from data creators to users and then connecting them at different levels and scales of online environments. The human mediators' efforts are usually invisible and easily overlooked compared to the technology mediators.

Therefore, when investigating the data sharing practices, this dissertation focused more on the work of human mediators than on technology mediators.

EOL and CyberSEES are two real-world cases chosen in this study. EOL is not only an aggregator repository, but also a global community of scientists, educators, students, nature enthusiasts, and staff from both research and public communities. CyberSEES is a cyberinfrastructure development project that uses Biocubes, a citizen science project, as a vehicle for collecting citizen science data. Biocubes encourages non-professionals, primarily educators and students, to observe and collect biodiversity data contained in one cubic foot space and share the data with the researchers and the world using the iNaturalist platform. iNaturalist is a global social network of naturalists and is not only a community of individual naturalists, but also a community of citizen science projects which adopt it as a data management tool and platform to form sub-communities of project organizers, data managers, and participants by creating project pages.

The findings about who the data providers and data mediators are revealed the data sharing contexts. This dissertation found that in EOL, there are seven types of collective-level data providers: venerable organizations, professional repositories, citizen science initiatives, social media platforms, education communities, subsidiaries, and academic papers, reflecting the general types of data sources in a certain knowledge domain (e.g.,

biodiversity). In CyberSEES, the (citizen science) data providers include Biocubes participants as individual-level data providers, the Biocubes project as the first collective-level data provider, and iNaturalist as the second collective-level data provider.

For EOL, the human mediators exist on both the EOL and the collective-level data provider sides. Core human mediators include the members of the two EOL working groups (i.e., SPGers and BIGers), EOL contractors, and the data creators/authors, data managers, and technicians from the collective-level data providers. For CyberSEES, the data mediators include Biocubes organizers (i.e., CyberSEES project members), iNaturalist data managers and technicians, iNaturalist community members, and EOL SPGers, BIGers, and contractor developers.

The data sharing processes are not static, one-time, or once-for-all. Instead, the data sharing processes are dynamic and sustainable collaborative efforts made by human actors with the support of technology. The core human mediators drove and participated in these collaborative efforts. The processes vary from one data provider to another. Each data provider is influenced by its organizational identities activated by the culture and institutional logics in which it is embedded. Without the help from human mediators, data sharing across research and public communities would be impossible. To summarize, in order to share data effectively in online environments across communities, the processes should include the following collaborative efforts:

- Data contributors generate data and upload data to home repositories in various online environments;

- Individual human actors from the research and public communities, who have enough domain knowledge, review the data and turn it into shareable and trustworthy research-grade data with appropriate licenses;
- Data managers and technicians from home and aggregator repositories build reliable data sharing partnerships and develop technology data sharing connectors to transfer data from home repositories to aggregator repositories;
- Data contributors, data managers, and technicians maintain data sharing partnerships and keep data and data sharing connectors updated.

Based on the understanding of the data sharing practices (i.e., data sharing contexts and processes) in these two cases, a new integrated framework of research data sharing was developed, illustrating the ecosystem of the knowledge infrastructure of data sharing across research and public communities.

## 7.1 Contributions

This dissertation addresses a critical but under studied aspect of the entire system of publicly funded science: promoting a truly inclusionary and democratic approach to science by sharing data effectively across research and public communities. Soranno et al. (2015) pointed out the central role of sharing data with the public in a feedback loop underlies the system of publicly funded science. However, there are limited previous studies investigating the collaborative efforts made by human actors, with the support of technology, on enabling data sharing from the research to the public community. In other words, it has been realized and confirmed by many scientists from different research

domains, as well as by society as a whole, that sharing data with the public is important (e.g., Borgman, 2015); however, nobody has yet explicitly explained how to achieve this. This dissertation addresses this issue by investigating two real-world case to reveal how data are shared effectively, by whom, through what processes, and in what contexts.

The specific contributions of this dissertation include: 1) filling the knowledge gap about research data sharing and promoting data sharing culture by exploring how research data can be shared beyond the research community; 2) extending research on data sharing across research and public communities from both horizontal (i.e., breadth) and vertical perspectives (i.e., depth); 3) making the invisible part of infrastructure (i.e., human mediators) who specifically work on data sharing more visible; and 4) refining and developing a data sharing framework and providing insight on data sharing practices and system design for sharing data across research and public communities. In addition, this dissertation not only contributes to data sharing practices across communities, but also has boarder contributions to data sharing practices in general. The four specific contributions are described below.

First, this dissertation contributes to fill the knowledge gap about research data sharing. Previous studies on research data sharing have been predominantly focused on sharing data created by researchers within the research community (Kowalczyk & Shankar, 2011). However, the phenomena that sharing research data openly (including with non-professionals from the public) and that research data could be created by non-professionals have become more common in recent years. This dissertation extended the

study of research data sharing communities from within the research community to across research and public communities by studying data sharing practices in 1) an online environment that supports research data shared not only with the research community, but also with the public community; and 2) an in-depth data sharing journey of the research data collected by non-professionals.

Second, this dissertation extends the research on data sharing across research and public communities from both horizontal (i.e., breadth) and vertical (i.e., depth) perspectives. From the horizontal perspective, this dissertation not only focused on data sharing practices of a single provider or one type of data provider but instead studied data sharing practices across communities in different types of collective-level data providers (i.e., professional repositories, venerable organizations, citizen science initiatives, social media platforms), for a total of over 30 different data providers. Each data provider represents a data source, a data infrastructure, and an organizational structure behind the data source and data infrastructure, indicating that the actual data sharing practices vary from one data source/provider to another. The findings in this dissertation shed light on the understanding of the diverse nature of collective-level data providers and the influence of this diversity on data sharing practices across communities. Extending the data sharing study from the horizontal perspective makes up for the insufficiencies of previous studies that only focus on one data source or one data sharing platform without the opportunity to compare different types of data sources and platforms.

From the vertical perspective, this dissertation is the first in-depth study on data sharing that observed and tracked data sharing practices from the data being originated in an offline environment to being shared and mobilized on different platforms across research and public communities in online environments (i.e., from a home repository to an aggregator repository). Data sharing practices included both sharing contexts and sharing processes. This kind of research design allows a deep and comprehensive understanding of the human actors' collaboration and technology support within the ecosystem of data sharing. Extending the data sharing study from the vertical perspective makes up for the insufficiencies of previous studies that mainly focus on one specific stage of data sharing (e.g., sharing to a home repository) or do not explicitly specify what stage of data sharing (e.g., general sharing) is studied. These two types of previous studies have not investigated what exactly happens to data during different stages of sharing. Sharing research data in the large scale online environment should not be static one-step sharing, but dynamic and sustainable multi-stage sharing.

The third contribution of this dissertation is that it was one of the first in a group of studies that made an effort to make the invisible part of infrastructure become visible. In previous studies on data sharing, the work of human mediators is usually invisible and overlooked (Kervin, Cook, & Michener, 2014; Borgman, 2015). However, without the human mediators' efforts, the technology infrastructure would not be developed and filled with meaningful content (Edward et al., 2013). The content focused on in this dissertation is biodiversity data. Previous studies have realized the inevitable sociological challenges, and agreed that solving these challenges is even more urgent

than solving the technology challenges of data sharing (e.g., Parr, Guralnick, Cellinese, & Page, 2012). This dissertation focused on revealing the human actors' work, especially collaborations between different parties (e.g., EOL working group members vs. content partner data managers, technicians, and data contributors; Biocubes organizers, participants, and iNaturalist community members). In order to further develop data sharing technologies and promoting data sharing culture, it is important to better understand the current practices of human actors, especially practices of core human mediators, to make data sharing take place at different levels (i.e., the collective and individual levels; on the home and aggregator repositories). Through this increased understanding, designers and developers can gain better insights on developing and designing next generation technology infrastructure to support human actors share research data not only within the research community, but also with other non-research communities.

Last but not least, this dissertation contributes by developing a new integrated framework of research data sharing. This framework extends the scope of the previous data sharing framework and illustrates the ecosystem of data sharing across communities, as well as the components of the knowledge infrastructure that include both technology and human infrastructure. Firstly, this framework does not limit data sharing within the research community; second, it includes the macro- and micro-influential contexts of sharing (i.e., collective level and individual level); third, makes data mediators as visible as technology mediators; and, forth, points out the mutually reinforcing relationships between data sharing practices and data sharing culture. This framework can be used to guide

researchers, practitioners, and system developers and designers to understand data sharing contexts and processes.

## 7.2 Limitations

The limitations of this dissertation predominantly stem from the research method. The case study approach enables researchers to do an in-depth investigation of cases (Hyett et al., 2014), as this dissertation did. By focusing on two cases, this dissertation investigated in-depth every step of data sharing processes from the data being created and shared in a home repository (i.e., CyberSEES) to being transferred to and shared with a large scale aggregator repository (i.e., EOL). However, the two cases could not be considered representative of and generalizable enough to all other cases of sharing data across research and public communities.

EOL is a large-scale aggregator repository and supports collective-level data providers in sharing data on its own platform, but it cannot represent all other aggregator repositories. Table 6.2 shows that EOL has strong support from different types of institutions and abundant funding from more than one foundation. Also, since the author of this dissertation started to collect research data from this case (2014), it has launched the second version of site. The development of the knowledge infrastructure, including both human infrastructure and technology infrastructure, has reached a level of maturity. Therefore, the findings of this case could be more applicable to the aggregator repositories at a similar large-scale and with strong institutional and financial support

than those than those who aggregate smaller-scale data, do not have such powerful institutional and/or financial support, and/or are still at an early stage of development.

For example, in the case of EOL, data are aggregated from seven types of collect-level data providers, which represents a great amount of time and energy by human mediators. Aggregator repositories with less institutional and financial support might not have these diverse types of data providers; instead, they might focus on aggregating data from only one or limited types of data providers that are most accessible and available to them.

Unlike the human mediators of EOL that are located in two independent organizations (i.e., one focuses on the contents of data, another focuses on technology issues) with help from a small group of contractors, a data aggregator repository with less institutional and financial support might have fewer human actors who all belong to the same organization. For technology data mediators, they might not have developed the infrastructure to match all the functions of EOL, but instead focuses on a displaying or exporting function.

For data sharing processes, this dissertation found that no matter what scale and what level of support a data aggregator repository might have, it requires a significant amount of time from human mediators to build reliable and trustworthy data sharing partnerships. It is unlikely an authoritative data source would allow any data aggregators to republish their data in any way without gaining approval or agreement from the persons in charge of these data sources. For data aggregator repositories with less institutional and financial support, it might take their human mediators greater time and energy to gain

trust from the data sources to build a partnership with them and obtain their data. With less strong institutional support, a data aggregator repository might not attract as many data sources as EOL.

Biocubes, developed by CyberSEES, is a small-scale citizen science project that supports individual-level data providers in creating and sharing citizen science data on the iNaturalist platform. Table 6.4 shows the number of individual-level data providers for Biocubes studied in this dissertation is fewer than 30. It cannot represent all other citizen science projects and other data sources created by non-professionals. It also has limited institutional and financial support and does not have enough human and financial resources to build its own data collection and management tools and repositories. Therefore, it adopted iNaturalist as the data collection and management tool and repository.

There are many citizen science projects with low to moderate institutional and financial support similar to Biocubes; likewise, they are not able to build their own technology infrastructure and rely on existing tools and platforms built by a third party for supporting citizen science. Most citizen science projects which adopt iNaturalist as the data management or even project management tool are this type of small-scale citizen science project. Until now, there are more than 6,000 project pages created on iNaturalist. However, as citizen science has been developing rapidly in the past few decades, there are also many citizen science projects with strong institutional and financial support and many of them are able to build their own information system for collecting and managing

data. In addition, as crowdsourcing technology has been adopted in citizen science, there are several large-scale successful citizen science projects with tens of thousands of individual-level data providers (e.g., eBird, Galaxy Zoo). Therefore, Biocubes cannot represent all these citizen science projects.

There exists many citizen science projects, built at different scales and in different forms. These projects could have any number of individual-level data providers from diverse communities of the general public; they could have simple or complex structures of project organizers; they could have their own data collection and management tools and repositories or adopted existing ones; and their data sharing processes—from collecting to sharing data to the choice of home repository—could be very different from each other. The findings from CyberSEES regarding data providers, project organizers as the human mediators, and data sharing processes from the individual-level data providers to the home repositories could not apply to different citizen science projects in these various conditions.

However, despite the limitations introduced above, this dissertation still provides significant value to all data aggregators and citizen data sources who are making an effort to share their data across research and public communities. For data aggregator repositories, on the one hand, the findings in the first case showed the possibility of building data sharing partnerships with different types of data providers; on the other hand, the findings illuminated data sharing processes and that they vary among specific data providers. No matter what the scale of data aggregator repository and how much

institutional and financial supports it has, the findings in the first case could help it to 1) evaluate its own resources and abilities; 2) decide which types of data provider to approach; and 3) think ahead about how to achieve agreement, set up data connectors, and make what kind of maintenance plan with them. By applying the findings, a data aggregator repository would be able to develop the most efficient and effective way of aggregating data from data providers.

For citizen science data sources, no matter whether the sources are created by citizen science projects or on social media platforms, data sharing processes from home repositories to aggregator repositories (i.e., building a formal data sharing partnership, reviewing data and increasing data quality) should be encouraged and promoted to be applied to all citizen data sources. These processes showed an effective strategy to ensure the reliability and trustworthiness of data created by non-professionals for the purpose of sharing it with both researchers in research community and non-professionals in the public community.

### 7.3 Future papers

Future work will first focus on translating the major topics and important findings of this dissertation to a series of publications. This section will introduce three potential papers with an overarching topic: effectively sharing research-level data through publicly accessible data sets across research and public communities in the context of infrastructure. These papers will address various research questions under this overarching topic.

The draft title of the first paper is *Infrastructuring Process of a Universal Global Scientific Data Hub*. This paper will be developed based on EOL. Infrastructuring is the process of developing and maintaining infrastructure (Karasti & Baker, 2004), which is an accurate and short term that could be used to refer the data sharing processes revealed in this dissertation. This paper will be in the form of a case study investigating the development of a knowledge infrastructure for aggregating and sharing biodiversity data across research and the public communities. The development stage focuses on the development of sociotechnical relationships (i.e., data sharing partnerships) between EOL and the diverse collective-level data providers after the fundamental technology infrastructure had been established. Without holding meaningful content (e.g., concrete data, information, knowledge), the technology infrastructure could not function by itself. The analysis focuses on revealing the human infrastructure from both the EOL and data providers' sides, the collaboration efforts made by the human mediators during the infrastructuring process, and the significant influence of the human mediators' available identities (from the data providers' side) on the specific infrastructuring process.

Besides the discussion conducted in this dissertation, this paper will discuss collaborative efforts directly made or driven by human actors (i.e., the human mediators) that could not be replaced by a machine during the infrastructuring process, most likely including the certain types of effort made at the early stage of building a sociotechnical relationship (e.g., achieving the agreement to aggregate data) and at the stage of maintaining this relationship (e.g., updating the data connector).

This paper will also discuss extending of the concept and definition of cyberinfrastructure. Based on previous concept of cyberinfrastructure, the EOL platform is not a product of cyberinfrastructure in a strict sense. Or in other words, EOL does not develop a knowledge infrastructure that can be considered a cyberinfrastructure if the concept or definition of cyberinfrastructure is bound to only or predominately serve the research community. Because EOL's major goals include sharing data with and for non-professionals and the public community, this paper will argue to extend the concept and definition of cyberinfrastructure to any knowledge infrastructure built not only for the research community, but also for the public community. This extension could promote scientific data sharing culture to include data sharing beyond the research community to the public community, especially given they are a stakeholder community in public funded research (Soranno et al., 2015).

The draft title of the second paper is *Standing on the Shoulders of Giants: Data Sharing Practices in a Small-Scale Citizen Science Project*. This paper will be developed based on Biocubes. The limitation section admitted that the Biocubes project is not representative of all citizen science projects and therefore could not be generalized to all citizen science projects. However, it can be considered a good representative of small-scale citizen science projects or the citizen science projects that have limited resources (e.g., human, technology, funding, etc.). A significant characteristic of these citizen science projects is their lack of resources does not allow them to develop a custom data infrastructure for their projects. However, sharing the data created by non-professionals

with scientists is a central goal of running a citizen science project. If scientists could not access and use the data, developing a citizen science project could be meaningless to the research community. Therefore, citizen science projects have to achieve their data sharing goals by choosing and adopting well-established data infrastructures that allow them to share their data with researchers and the research community. The *Giants* in the title of this paper refers to this kind of data infrastructure.

This paper will be a case study that investigates the data sharing practices of Biocubes. This investigation focuses on identifying data sharing practices and how the resources provided by CyberSEES influences these data sharing practices. Resources are transferred to and reflected by the human and technology infrastructure available in CyberSEES. The human infrastructure includes CyberSEES project members and Biocubes project participants. The technology infrastructure focuses on the well-established data infrastructure adopted by CyberSEES, iNaturalist. This paper will focus on discussing how the existing data infrastructure enables and restricts data collection and sharing Biocubes.

The third paper will be a theory paper about developing a framework of data sharing in the context of infrastructure. This paper will be developed based on the integrated framework of data sharing introduced in Section 6.3. There have been multiple data sharing frameworks developed by previous research which help us to understand the data sharing environment and scientific data sharing practices from different perspectives. However, a framework that could explain a comprehensive data sharing ecosystem (i.e., a

knowledge infrastructure) (Edwards, 2010) and reflect an inclusive data sharing culture (Soranno et al., 2015) is still missing. This framework should consist of human and technology infrastructure (i.e., the complex sociotechnical relationships) and reflect the data sharing process. The integrated framework of data sharing introduced in this dissertation can be considered a draft version of this missing framework.

The current version of the integrated framework was developed based on four existing data sharing frameworks and the in-depth investigation of the data sharing practices of the two cases in this dissertation. However, it is still too complicated, and there is room for improvement. After modifying and simplifying the current version, this theory paper will introduce an updated version of integrated framework. This paper will also examine whether the updated version of the framework could be applied to another two or three real-life data sharing cases in which the data are successfully shared across research and public communities.

#### 7.4 Future work

The future work will then focus on extending the depth and breadth of the understanding of data sharing practices across research and public communities.

For extending the depth of the study, future work will add a quantitative method to leverage the current qualitative method and will investigate the outcomes of data sharing enabled by established partnerships. The outcomes should reflect the results of the collaborative effort between the human mediators.

Take the first case in this dissertation, EOL, as an example to illustrate how the next step will look. The results of data sharing can be reflected by the archived and publicly viewable usage statistics records for each content partner. The Usage Statistics is a function that EOL infrastructure provides for recording statistics concerning each content partner at a monthly frequency (Figure 7.1). The statistics results for each of these items are published each month after the partnership has been successfully built between EOL and a content partner. The column titled “All pages” refers to the usage recorded for all the species pages that EOL currently has. The column titled “Provider percentage” is the usage statistics for the pages contributed by each content partner. In addition, EOL also provides usage statistics for each single species page that contains data shared by a content partners.

Statistic measured for the month of March, 2015	Pages containing provider content	All pages	Provider percentage
Pages (number of pages)	5007	4668861	0.11%
Pages viewed (number of pages viewed by visitors)	102	212592	0.05%
Unique page views (number of visitors viewing pages)	199	1423832	0.01%
Page views (number of visits to pages)	234	1815990	0.01%
Total time spent on pages (hours)	2.95	8657.0	0.03%

Figure 7.1 Example of the summary of the usages statistics for a content partner.

Based on the usage statistics, a data sharing pattern for each content partner can be identified. The next step of future study will start by identifying the data sharing patterns for the content partners whose JIRA system contents have been analyzed qualitatively in this dissertation. Combining the qualitative analysis results and the quantitative data sharing patterns will reveal how collaboration between the human mediators influences

data sharing patterns. The usage statistics adds another type of important data source to study whether the data sharing pattern also varies in the same way or different ways.

For extending the breadth of the study, the next step will be to expand the scope of the origin of data. This dissertation chose Biocubes as the first scenario to understand how and why data created by non-professionals were originated and later shared in online environments. iNaturalist's data entry system and data exhibition system are particularly structured for its members to contribute data in a format similar to traditional research data (i.e., biodiversity occurrence data). This platform values the importance of citizen science and promotes the development of it throughout the entire platform. Future research will study data created by non-professionals and originated in more casual offline environments and shared through platforms like Twitter and Flickr. This kind of platform are not structured for users to contribute data in a research data format.

Furthermore, the scenarios in which data contributors collect data are also less likely to be citizen science projects with explicit introductions and protocols for collecting data for research purposes.

Daume and Galaz's study (2016) have reported that some Twitter users used real-time information about biodiversity observations with species determination requests. Twitter users received feedback from other Twitter users' replying to the identification request. These data become valuable research data (i.e., ecology monitoring data) in a process similar to iNaturalist. Therefore, future research will first repeat and adapt Daume and Galaz's study to investigate what collaboration efforts are made by Twitter community

members as human mediators to identify the species, how the differences in infrastructure design (i.e., iNaturalist vs. Twitter) could influence collaboration efforts, and after Twitter data (i.e., information about real-time biodiversity observation) become research-grade data, how they would be shared and mobilized onto different platforms.

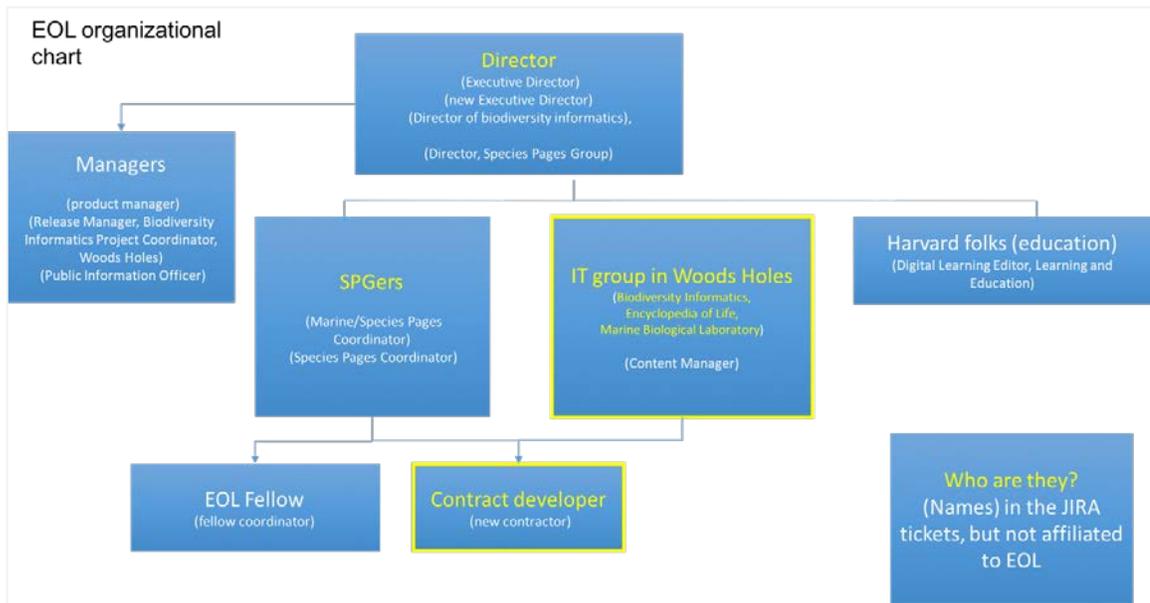
To sum up, this chapter summarized the core findings of the two in-depth cases about data sharing practices for research data across research and public communities. Then the limitations rooted in the research method and design were introduced. At the end, future papers and future work directions were introduced. This dissertation is not an end point, but the start of investigating research data sharing and developing a data sharing culture beyond the traditional research community.

# Appendices

## Appendix One – Interview slides for the case of EOL

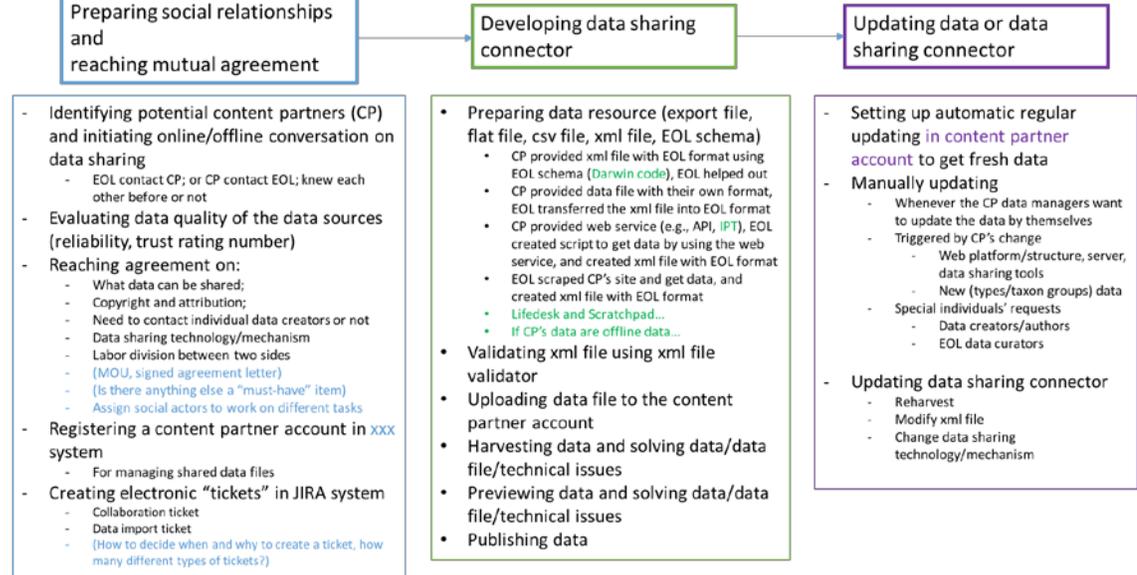
For the first interview:

The slide for verifying EOL organization structure and human mediators' organizational identities:



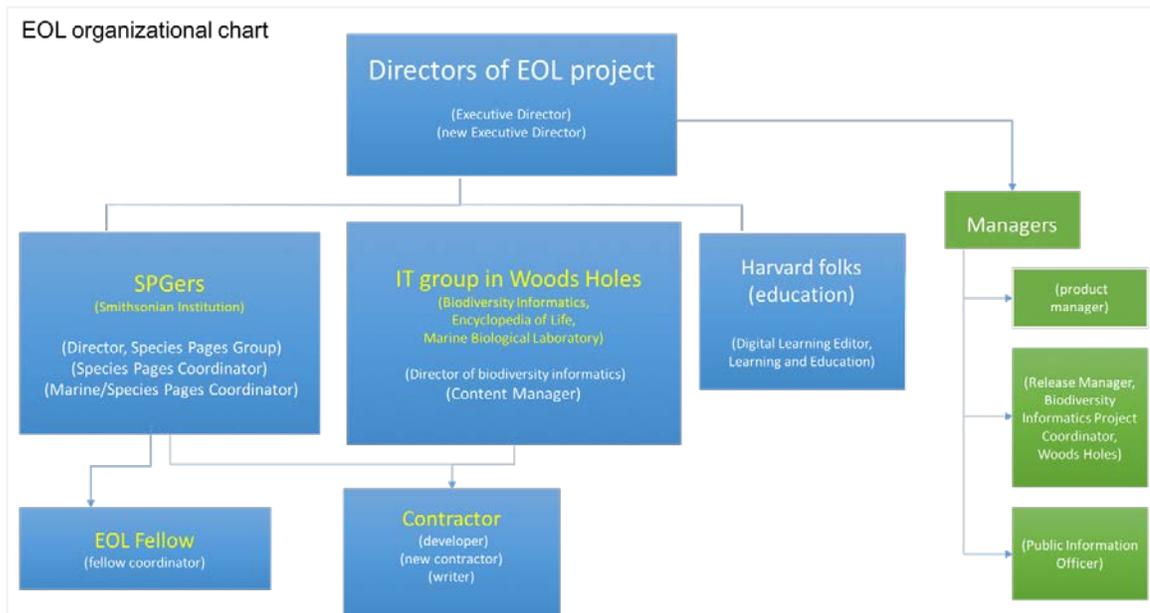
The slide for verifying the data sharing processes:

EOL data sharing processes = formation of partnership and maintenance of partnership



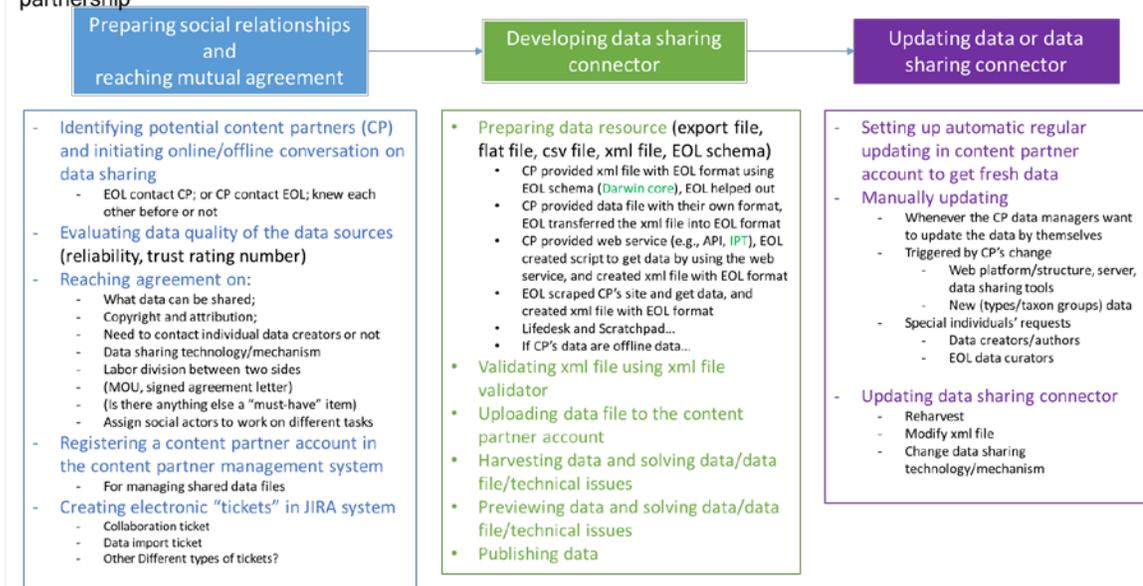
For the second interview:

The slide for verifying EOL organization structure and human mediators' organizational identities:



The slide for verifying the data sharing processes:

EOL data sharing processes = formation of partnership and maintenance of partnership



Appendix Two – Initial coding schema

1. For when to share data? Review content partners’ own websites and databases to figure out where their data came from, and at what stage of data life cycle (Michener & Jones, 2012; Rüegg et al., 2014).

Plan  
 Collect  
 QA/QC  
 Analyze  
 Descrone  
 Preserve publish  
 Discover  
 Integrate  
 Analyze

or

Plan  
 Collect  
 Assure  
 Describe  
 Preserve  
 Discover  
 Integrate  
 Analyze

2. For what components are included in data sharing? the coding start list can be created based on the data sharing framework (Fecher et al., 2015):

<b>DP</b>	<b>Anything about data provider/donor from individual level</b>
DPSF	Sociodemographic factors
DPDC	Degree of control (e.g., how the CPS [content partner staff] who is the contact person of CP can control the process of setting up the partnership)
DPRN	Resources needed (e.g., what resources do the CPS need to set up partnership with EOL)
DPRT	What does the CP can get out from setting up partnership with EOL, what returns to them by having this partnership with EOL.

<b>DPO</b>	<b>Anything about data provider/donor's organization</b>
DPOFA	funding agencies

<b>DPC</b>	<b>Anything about data provider/donor's community (e.g., research community)</b>
DPCDSC	Data sharing culture
DPCDST	Data sharing standards
DPCDSV	Data sharing value
DPCPB	Publications, the primary currency in academia

<b>NM</b>	<b>Norms</b>
NM-E	Ethical norms
NM-L	Legal norms (e.g., copyright)

<b>DI</b>	<b>Anything about data infrastructure (i.e., technology infrastructure)</b>
DIAT	Architecture
DIUB	Usability
DIMT	Management software and tools
DIHW	Human workers

<b>DR</b>	<b>Anything about data recipients/consumers/users</b>
DRU	Data usages for what, can be both appropriate use and adverse use
DRO	Data recipient's' organization

DRSF	Sociodemographic factors
------	--------------------------

3. For what culture and institutional logics are the components embedded in, the coding start list can be created based on institutional logics theory (Thornton et al., 2012):

<b>SI</b>	<b>Social interaction between social actors</b>
SI-Within	Social interaction contents within core project team, such as between ES and ESC [EOL staff who are in charge of setting up partnership between EOL and a CP], or such as between Biocubes project members.
SI-Between	Social interaction between core project members with data providers or data recipients, such as between ES [EOL staff] and CPS (Content partner staff), and between Biocubes project members and data providers/users (partners).
SI-DM	Decision making (e.g., what decision is made by who; what is the process of making the decision)
SI-SM	Sensemaking (e.g., qualification for the issues and actions)
SI-Mobilization	Group level motivation

4. For what do the components stand for? the coding start list are created based on organizational identification theory (Ashforth et al, 2008):

- Core of identity;
- Content of identity;
- Behaviors of identity.

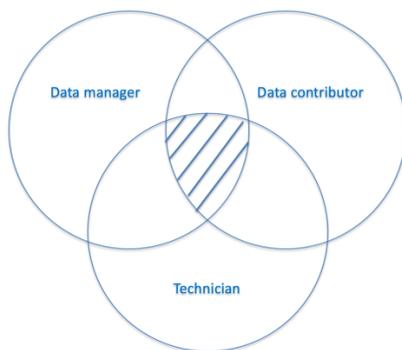
<b>OI</b>	<b>Organizational identities</b>
OIDP = DPSF	Organizational identities for data provider (i.e., CP), for the Biocubes project, for project members.
OIDP>Core	Core of identities
OIDP>Content	Content of identities
OIDP>Behaviors	Behaviors of identities
OICPM	Organizational identities for ES [EOL staff]/ BM [Biocubes project members]

FOA	Focus of attention, for example priority of certain type of tasks.
OICPM>Core	I am "A" (self-definition), I value "A" (importance), I feel about "A" (affect)
OICPM>Content	I care about "B" (values), I want "C" (goals), I believe "D" (beliefs), I generally do "E" (Stereotypic traits), I can do "F" (Knowledge, skills, abilities)
OICPM>Behaviors	I do "G" (behaviors)

### Appendix Three: The data sharing stories

In the following, four true stories represented four scenarios will be introduced to illustrate how the processes of data sharing vary from case by case and why “human” is the most important and significant influential factor on the data sharing processes. The four stories are true data sharing stories of four EOL content partners.

#### **The first story**



The data manager, data contributor, and technician is the same person. This story matches the cell # 1 in Table 4.1.

The first story is about a professional repository data provider whose data manager, data creator, and technician are the same person. It is a data sharing story happened between EOL as an organization and an individual data owner. This data source contains large number of images and text data objects of multiple groups of organisms. This content partner’s own website is a web 1.0 site and developed by the data manager himself for mainly share his data, but also a very small amount of data from his colleagues and friends.

The data manager and EOL had existing relationships before EOL human mediators invited this data manager to become a content partner. When the early version of EOL technology infrastructure was still under development, this data manager had provided a few data objects (i.e., images). EOL data mediators manually uploaded these data objects and displayed them on EOL's exemplar pages. In addition, a university professor had suggested EOL to invite this data manager to become a content partner.

After the first version of EOL technology infrastructure is developed, EOL turned its focus of attention from developing infrastructure to developing the community of biodiversity data partners, in other word, developing the partnerships with diverse biodiversity data providers. EOL human mediators then sent the invitation to the data manager via email to ask the possibility of building formal collaborative data sharing partnership (March, 2008).

At the end of the 2008 around November, the data manager accepted the invitation and agreed to build the partnership with EOL. He pledged to convert the copyright of his website to the creative commons, and provided attribution information that he would like to display on EOL for his data.

#### *Step 1: Reaching agreement*

The data manager's conversation about how to proceed the collaboration started with three administrative human actors. The EOL content manager first helped confirm with the data manager about his agreement on building the partnership with EOL. Then the

data manager handed over this conversation to the members of EOL Species Page Working Group (SPGers). In the 2010, an EOL SPGer, a species page coordinator, found the license conversion of this data provider's website has been done, and would like to get back in touch with the data manager. The EOL SPG director agreed and assigned this EOL SPGer to manage the partnership building with this data manager.

The EOL SPGer and the data manager first discussed the copyright of data and confirmed what license should be used for the data objects. They then discussed what data the data manager would like to share. The data manager agreed to share both image and text type of data. Then they discussed the data sharing mechanism, asking the data manager's preference on using what method to share data. The data manager chose to provide a flat-file, rather than a xml file in EOL schema, to allow EOL technicians to scrape his website. Also they discussed the updating frequency. The data manager pointed out that he usually rebuilds his website twice a year, and suggest EOL to do the updating of his data on EOL at the same time. The data manager empathized that the updating his data on EOL is very important for him and wanted to sorted this out early on.

### *Step 2: Building data connector*

After the data manager and the EOL SPGer reached the consensus on the basic agreement items mentioned above, the data manager prepared the flat-file and sent it to EOL SPGer. The EOL SPGer then forwarded this flat-file to the EOL technician. The EOL technician prepared the data resource by generating a XML data file in EOL schema from the scraping the data provider's website. The EOL technician also created a content partner

account and upload the data resource into the account for the data provider. Till now, the data sharing connector was created successfully.

The EOL technician started the harvest process by running the data sharing connector. After the data has been harvested, the EOL SPGer reviewed the data carefully on the preview mode, and figuring out whether there are any problems on the preview mode. The EOL SPGer then found some problems (e.g., names displayed incorrectly, empty pages) which then was fixed by the EOL technician. The EOL technician found asking more technological information about the data from the data manager could improve the data sharing. The technician asked the EOL SPGer's help to ask the data manager to provide more information about his data. After the EOL SPGer got the information from the data manager and sent it to the EOL technician. The EOL technician improved the data resources and then run the data connector and harvested the data again.

After the display of data looked good for the EOL SPGer, the data manager was invited to review the data on the preview mode. The data manager made a small suggestion on how to display the text data, and then approved the data preview. The data manager also provided an exemption from the speed restriction to allow EOL's to get data faster in the next time harvest.

After the EOL SPGer got the approval from the data manager on data preview and publish data, a time is scheduled for formally publish the data on EOL. The data were then officially published on EOL for the first time (September 2010), representing the

successful formation of the partnership. The data provider officially became a EOL content partner.

*Step 3: Updating data connector*

Building the partnership with EOL makes the data provider started to consider how to improve the management of his data. He told the EOL SPGer that the opportunity of building the partnership with EOL motivated him to make time to think about the current condition of how he store and share data and how can the current condition can be improved in the near future.

Later on, the data manager decided to divided his website into two sites for storing and sharing different types of data. Then he decided to change the method of sharing data with EOL. He contacted the EOL SPGer about the changes he would like to make for his websites, and would like to choose to provide multiple CVS files of different types of data to EOL, so that EOL did not have to use flat-file to scrape the website anymore.

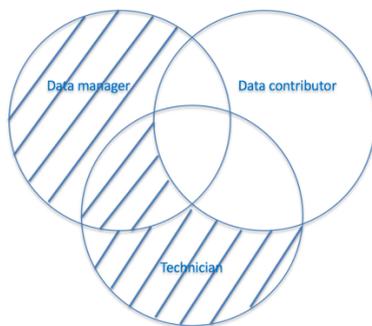
After the website separation settled down, the data manager provided the EOL SPGer the CVS files and instructed the EOL SPGer to understand the contents in the CVS files. After the EOL SPGer review the CVS files, the data managers and the EOL SPGer discussed what content can be improved in the CVS files and finished preparing the CVS files. Then EOL technician helped to prepare the data resources based on the CVS files and updated the data sharing connector. After the data have been successfully reharvested, the data manager, the EOL SPGer, and the EOL technician repeated the

processes of data previewing and problem solving again, and published the latest version of data.

In about two years, the data managers made another changes by transferring his image data to a new web sever. He decided to asked EOL's collaboration again to change the data sharing method from generating CVS files to sharing a metadata file. Sharing metadata file would be able to allowed the EOL technician to generate the DWC-A file as the new type of data resources.

Step by step, the collaboration between the data provider and EOL led to a much stronger and more stable data sharing connector. In turn, it represents the partnership between this data provider and EOL became better and better over time.

### **The second story**



Data managers and technicians is the same person. This story matches cell #2 in Table 4.1.

The second story is about a venerable organization data provider who has multiple different level data managers (multiple administrative social actors) and technicians. Some data managers themselves are also technicians (Figure 2).

*Step 1: Reaching agreement*

The data sharing conversation started between the director of EOL and the administrative human actors from this venerable organization data provider. The director of EOL visited the organization to discuss the potential collaboration opportunities. Unlike the previous data sharing story happened between EOL as an organization and the individual data owner, this story happened between EOL as an organization and another organization as data provider.

The administrative human actors from the data provider showed interests in developing collaborations. They hoped the collaborations were not limited to just sharing data, but also including getting the scientists working for the data provider involved in EOL community, and joined funding opportunities. A formal agreement letter was needed and should be agreed and signed by both sides' directors.

After the data provider granted the oral agreement on building the collaborations and clarified what types of data can be shared with EOL, while the formal agreement letter was under development, the administrative human actors from the data provider and the EOL directors handed over this conversation to a small group of technicians from the data provider and EOL SPGers. One of the technicians was the core data manager who

could represent the data provider and coordinate the work of building the partnership with EOL.

*Step 2 & 3: Building data connector and updating connector*

The core data manager who was also a technician from the data provider first chose to provide a CVS file to EOL SPGers and manually update the data by updating the CVS file. The EOL SPGers helped to set up a content partner account and prepared the data resource based on this CVS file. Later on, the technicians from the data provider and the EOL SPGers discussed the possibility of building closer data sharing collaboration by developing a more automatic way of sharing data, so that the data provider's data on EOL could get updated in a timely manner. This also increase the possibility of sharing more types of data.

On the one hand, the progress on sharing more types of data went slow. It is because the data provider is a venerable organization and holds professional and reliable data sources. Their data that would be shared and reposted on other website are closely associated to this organization's reputation. They need to ensure that the data are good enough and are very well prepared before they can make the decision of sharing. Therefore this organization were very careful about sharing what data. In addition, even though the data quality could be controlled, they had concerns about the possibility of their data being misrepresented or misused in different contexts. After the data are shared on EOL, how the data would be used is not under their control.

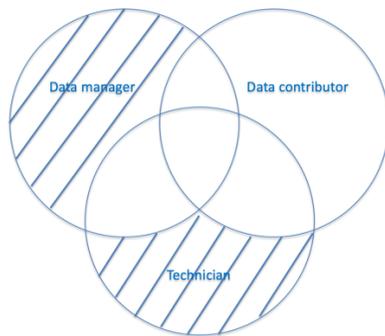
On the other hand, however, the progress on building a more automatic way of data sharing went smoothly and fast. It is because another group of technicians from the data provider had already worked on developing the API (application program interface) for a while and was looking for the first users of the API. The technicians from the data provider were interested in knowing who would like to use the API to retrieve what data in what ways from their data sources. So that they could better understand the data users' needs and improve the API. The core data manager who is also a technician from the data provider thought that EOL could be a ideal first user of this API. Later, the EOL SPEers agreed this core data manager's proposal, and decided to try their API. They thought it could be a better way of getting the data from the data provider, more efficient and less human effort was needed.

The technicians who developed the API from the data provider then helped EOL SPGers to learn what data the API could provide and how to access it. Then the data manager who is also a technician and the EOL SPGers discussed the copyright and license issues and confirmed what data were allowed to get from using their API. Therefore, not all data that the API could provide would be allowed to publish on EOL. The decisions of what data could be shared, what license of data should be used, and other terms and condition of using API were made after another round of discussion between the EOL SPGers and the administrative human actors from the data provider.

By following the introduction provided by the technician from the data provider, the EOL technician used the API to create a data resource and created a new data connector. When

the data were successfully harvested, the EOL SPGers proved the data preview first and invited the data manager from the data provider to review the data on the preview mode. The EOL SPGers guided the data manager from the data provider to use EOL content partner registrar system to review the data. Then the EOL SPGers, the EOL technician, and the data manager from the data provider did iterative process of solving data issues, and confirmed the attribution, license, and logo information that should be added when display the data provider's data on EOL species page. After finishing the data preview and receiving the approval from the data provider, the rest of the processes for publishing data are the same with other content partners.

### **The third story**



The data manager and technician are different persons. This story matches cell #5 in Table 4.1.

The third story is about building partnership with a social media platform data provider. The previous two stories are about sharing data created by the researchers and experts. This third story illustrates how to build a pipeline of general public data between the social media platform and EOL.

The EOL product manager did most initial work on preparing the relationship, including having face-to-face and telephone meetings with content manager from the data provider to discuss the possibility of building the collaborative sharing relationship. After the content manager from the data provider decide to build the partnership with EOL, the EOL product manager introduced the members from EOL Biodiversity Informatics group (BIGers) and EOL Species Page group (SPGers) into the conversation between the data provider and EOL.

*Step 1: Reaching agreement*

Before EOL reached out to this social media platform, it has become successful and popular among netizens in United State and even in the worldwide scale. There have been large amount of amazing contents generated by the users were uploaded on this social media platform. The web service and sharing kit has been developed for supporting sharing these user generated contents to outsiders.

However, although there are contents of interests for EOL among those existing contents, it is impossible to build a connector to share these existing contents immediately to EOL because of two reasons: 1) there was not appropriate tag (e.g., a scientific name) attached to the existing contents yet; and 2) both the EOL and the data provider could not be able to have a person or a group of people searching and finding appropriate content (e.g., texts, images, audios, videos) that could be shared on EOL. The content manager from the data provider and EOL therefore agreed that the first step of building the connector

should be creating a EOL group within the existing community that consisted of all current registered users on this social media platform. Then current and future community members would be encouraged to submit biodiversity contents to the EOL group.

The content manager from the data provider and EOL SPGers, BIGers, technicians then collaboratively figured out how to allow the community members to add tag to the contents and what are the license that should be used when submitting the contents. The EOL human actors need to effectively communicate these requirements with the community members by asking them to follow certain guidance for submitting the content.

The EOL technician then created a content partner account for the data provider. He also used this data provider's web service (i.e., API) to generate the data resources for the data that would be submitted to the EOL group on this social media platform. The content manager from the data provider agreed to do a "shout out" to the existing community about their data sharing partnership with EOL. The "shout out" would inform the community members that there is an opportunity of sharing their data not only on this social media platform, but also on EOL.

### *Step 2 & 3: Building data connector and updating connector*

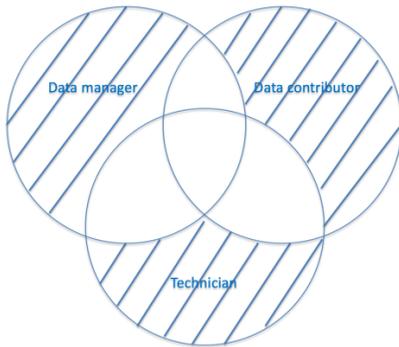
In order to build the data connector, the EOL project manager and technician created their own user account on this social media platform, and set up a EOL group. With the

help of a technician from the data provider, the EOL technician learned the following things: 1) how to use the web service to grab the information EOL wanted; 2) how to use machine tag to help community users to add tags to the content. He also was able to solve the technical problems encountered during the process of preparing and building the data connector.

In order to get the first wave of data, the EOL technician sent the invitation messages and guidance to several existing community users to encourage them to submit contents to the EOL group. The director of EOL SPG helped on commenting the invitation messages and guidance.

After the data had been successfully harvested, the EOL technician and the EOL SPGer reviewed the data in the preview mode. During the iterative process of reviewing data, the EOL BIGers helped the EOL technician to solve the technology problems about how to display a certain type of contents successfully on EOL page after the data have been harvested. The EOL technician later decided to re-harvest the data from this social media platform at weekly frequency.

## The fourth story



Data manager, data contributor, and technician are different persons. This story matches cell #6 in Table 4.1 in the dissertation thesis.

The fourth story is about a professional repository data provider. The fourth story is different from the first and second story from two aspects: 1) the data providers built their own websites and databases to manage their data in the in the first and second stories; and 2) the data contributors had not been involved in the partnership building in the first two stories. However, in the fourth story, the data provider did not build the websites or database by themselves, and they involved the data contributors in making the important decision of building the partnership with EOL.

At the beginning, a member from EOL species page group (SPGer) contacted the information manager in a venerable organization to explore the possibility of building new partnership between EOL and this venerable organization. This venerable organization have shared the data of one project with EOL in the past. The professional repository built for this project was hosted by this venerable organization.

The EOL SPGer hope to see whether it is possible to build connection with other projects and data sources within this venerable organization. There are many potential collective and individual level data providers in this organization. However, it hard for the information manager to figure out any other custodians of other data resources in this organization that could be pointed to EOL at the moment. She pointed out two major reasons: 1) most of them were under the pressure of having publication, and webpages were not usually considered as real publication; 2) a lot of data had not been well curated and not ready to share. This means that EOL had to put more effort by themselves on exploring and identifying potential data managers of different data sources within this organization.

*Step 1: Reaching agreement*

It took a while for EOL SPGers to finally find a data manager of a data source in this organization who were interested in sharing data with EOL. This data source does not have an independent self-service information system that was built just for sharing the data from this data sources. This data source was initially shared in an online environment by being uploaded in a professional aggregation type repository. So the data manager of this data source wanted to share the data to EOL via this aggregation type repository that was hosted by a different organization. The EOL SPGer first guided this data manager of this data source to set up a content partner account. They then talked about the license and attribution of the data, and confirmed that the rights statement of the data varies from some photographers to other photographers. So the permissions from certain individual data creators (i.e., photographers) are necessary.

They also discussed how to prepare the data resource file for building the data connector. This data manager felt it would be hard because the data live on another organization's server. So this data manager introduced the technician of the aggregation type repository to the EOL content working group, and hope they can figure out how to share the data from this repository to EOL.

The technician from the aggregator repository explained the current web service they could provide, however, the web service did not output the data document in the EOL format. The technician need to take time to work on understanding EOL's needs and write a script to transfer the data document in the EOL format. It would be too time consuming and complicated. This technician probably can only provide a one-time data dump of the data metadata. Then she would let EOL technician to create the data resource file that EOL needs by himself.

At this moment, there is another project that has uploaded their data in the same aggregator repository. But this project did not belong to neither the previous venerable organization that EOL hope to explore to build new relationship with, nor the organization hosted the aggregator repository. This project showed great interests in sharing their data with EOL via this aggregator repository.

The director of this project talked with both the EOL SPGers and the technician from the aggregator repository about providing a mini-grant to support the technician from the

aggregator repository to do two things: 1) help prepare the data resource file; and 2) set up an opt-in system by modifying the current aggregator repository system for enabling all the data contributors on the aggregator repository to share their data with EOL if they want. This opt-in system should also be able to provide a direct communication channel between the data creators and data users from EOL. For example, whenever there are data users from EOL to leave comments below the data shared by this project, these comments were not only sent to the collective-level email address of this project, but also directly routed to the data creators.

This project's strong motivation to share their data with EOL and their willingness to provide the financial support greatly facilitated the partnership establishment between EOL, the two projects, and the aggregator repository. One EOL content partner usually has one primary organizational identity. In this story, it is clear that there were three parties involved in building the partnership with EOL. They are all data providers. In the end, they decided to use the organizational identity of the content partner should be the aggregator repository. But the data sources from the two projects should be clearly displayed on EOL pages.

Later on, the EOL SPG, the two projects' data managers, and the technician from the aggregator repository discussed how many and what data could be shared. They weighed the workload of preparing the data resources documents for these data and setting up the opt-in system on the current aggregator repository. After they reached the agreement on

the work tasks and time, the technician from the aggregator repository signed a contract for working on building the connector between EOL and the aggregator repository.

*Step 2 & 3: Building data connector and updating connector*

By following the guidance and with the help from the EOL content working team and EOL technicians, the technician from the aggregator repository successfully built the data connector between the aggregator repository and EOL. The guidance from EOL technicians helped the technician from the aggregator repository learned each task in the processes of building the data connector and the underlying mechanisms of the data connector.

This technician also successfully set up the opt-in system on the aggregator repository. She also wrote a formal announcement about the partnership built between this aggregator repository and EOL, and posted it on its website. This announcement encourages individual data creators from this aggregator repository to share their data with EOL by using the opt-in system.

The processes of previewing the data preview and solving data problems were similar to the previous data sharing stories. After the data managers of the two projects got the permission of the individual data creators, the data were formally published on EOL.

## Bibliography

- Abeillia abeillei. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://www.eol.org/pages/1048780/overview>
- About Us. (2011). In *Biodiversity Information Standards (TDWG)*, Retrieved March, 2016, from <http://www.tdwg.org/about-tdwg/>
- Ackoff, R. L. (1989). From data to wisdom: Presidential address to ISGSR, June 1988. *Journal of applied systems analysis*, 16(1), 3-9.
- Acord, S. K., & Harley, D. (2012). Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0. *new media & society*, 1461444812465140.
- Albert, S., Ashforth, B. E., & Dutton, J. E. 2000. Organizational identity and identification: Charting new waters and building new bridges. *Academy of Management Review*, 25: 13-17.
- Alfred P. Sloan Foundation. (2016). Recently Completed Programs, Encyclopedia of Life. Retrieved March, 2016, from <http://www.sloan.org/major-program-areas/recently-completed-programs/encyclopedia-of-life/>

Atkins, D. E. (Chair) (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation.

Arzberger, P. W., Schroeder, P., Beaulieu, A., Bowker, G. C., Casey, K., Laaksonen, L., ... & Wouters, P. (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal*, 3(29), 135-152.

Ashforth, B. E. 2001. *Role transitions in organizational life: An identity-based perspective*. Mahwah, NJ: Erlbaum.

Ashforth, B. E., Harrison, S. H., & Corley, K. G. (2008). Identification in Organizations: An Examination of Four Fundamental Questions. *Journal of Management*, 34(3), 325-374.

Ashforth, B. E., & Mael, F. (1989). Social identity theory and the organization. *Academy of Management Review*, 14: 20-39.

Aves. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://www.eol.org/pages/695/overview>

Axelsson, A. S., & Schroeder, R. (2009). Making it Open and Keeping it Safe e-Enabled Data-Sharing in Sweden. *Acta sociologica*, 52(3), 213-226.

Bailey, C. A. (2007). *A guide to qualitative field research*. Sage Publications.

Barnard, C., & Simon, H. A. (1947). *Administrative behavior. A study of decision-making processes in administrative organization*. Macmillan, New York.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O., Swartz, B., Quental, T. B., ... & Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived?. *Nature*, 471(7336), 51-57.

Berman, F. (2001). The human side of cyberinfrastructure. *EnVision*, 17(2), 1.

Bettenhausen, K.L., & Murnighan, J. K. 1991. The development of an intragroup norm and the effects of interpersonal and structural challenges. *Administrative Science Quarterly*, 36: 20-35.

Bietz, M. J., Baumer, E. P., & Lee, C. P. (2010). Synergizing in cyberinfrastructure development. *Computer Supported Cooperative Work (CSCW)*, 19(3-4), 245-281.

Biocubes. (2015). In *iNaturalist*. Retrieved March, 2016, from

<http://www.inaturalist.org/projects/biocubes>

Biodiversity Informatics Working Group. (n.d.). In *Encyclopedia of Life*, Retrieved March, 2016, from [http://www.eol.org/info/biodiversity\\_informatics\\_working\\_group](http://www.eol.org/info/biodiversity_informatics_working_group)

Blaustein, R. (2009). The encyclopedia of life: describing species, unifying biology. *BioScience*, 59(7), 551-556.

<http://bioscience.oxfordjournals.org/content/59/7/551.full>

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 977-984.

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 1059-1078.

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational researcher*, 33(8), 3-15.

Boudreau, M.-C. and Robey, D. (1999). Organizational transition to enterprise resource planning systems: Theoretical choices for process research. In Proceedings of the 20th International Conference on Information Systems. Charlotte, NC.

Bowker, G. C. (1994). *Science on the run: Information management and industrial geophysics at schlumberger, 1920–1940*. Cambridge, MA: MIT Press.

Bowser, A., Wiggins, A., Shanley, L., Preece, J., & Henderson, S. (2014). Sharing data while protecting privacy in citizen science. *interactions*, 21(1), 70-73.

Brett, J. (2010). Clueless About Culture and Indirect Confrontation of Conflict. *Negotiation and Conflict Management Research*, 3(3), 169-178.

Brewer, Marilynn B. Krueger, Joachim I. (Ed), (2008). Depersonalized trust and ingroup cooperation. *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes. Modern pioneers in psychological science: An APS-Psychology Press series.*, (pp. 215-232). New York, NY, US: Psychology Press, vii, 391 pp.

Carlson J, Stowell-Bracke M. Data management and sharing from the perspective of graduate students: An examination of culture and practice at the Water Quality Field Station. *Libraries and the Academy*. 2013; 13: 343–361. doi: [10.1353/pla.2013.0034](https://doi.org/10.1353/pla.2013.0034)

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive psychology*, 17(4), 391-416.

Cohen, J.E. (2008). Privacy, Visibility, Transparency, and Exposure. *Univ. of Chicago Law Review*, 75(1).

Content Partners (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://www.eol.org/content\\_partners](http://www.eol.org/content_partners)

Cornell Lab of Ornithology. (2016). View and Explore Data. In *eBird*. Retrieved March, 2016, from <http://ebird.org/ebird/explore>

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). Managing and Sharing Research Data.

Creative Commons. (n.d.). Retrieved March, 2016, from <https://creativecommons.org/licenses/>

Crowston, K. (2000). Process as theory in information systems research. In *Organizational and social perspectives on information technology* (pp. 149-164). Springer US.

Crowston, K. (2000). Process as Theory in Information Systems Research. In R. Baskerville, J. Stage, and J. I. DeGross (Eds.), *Organizational and Social Perspectives on Information Technology* (pp. 149-164), Boston: Kluwer Academic Publishers.

Curry TE, Ansolabehere S, Herzog HJ (2007) A Survey of Public Attitudes Towards Climate Change and Climate Change Mitigation Technologies in the United States:

Analyses of 2006 Results (Laboratory for Energy and the Environment, Massachusetts Institute of Technology, Cambridge, MA).

Cyert, R. M., & March, J. G. (1963). A behavioral theory of the firm. *Englewood Cliffs, NJ*, 2.

CyberSEES project description document (2014). CyberSEES: Type 1: Collaborative Research: Infrastructure and Technology Supporting Citizen Science Data Usage and Distribution for Education and Sustainability.

Darwin Core. (2015). In *Biodiversity Information Standards (TDWG)*, Retrieved March, 2016, from <http://rs.tdwg.org/dwc/>

Darwin Core Text Guide. (2015). In *Biodiversity Information Standards (TDWG)*, Retrieved March, 2016, from <http://rs.tdwg.org/dwc/terms/guides/text/>

Daume, S., & Galaz, V. (2016). “Anyone Know What Species This Is?”—Twitter Conversations as Embryonic Citizen Science Communities. *PloS one*, 11(3), e0151387.

Deci, E. L., & Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.

drop. (2015, January). Atlantic Ribbed Mussel. In *iNaturalist*. Retrieved March, 2016,

from <http://www.inaturalist.org/observations/1191926>

drop. (2015, March). Image of *Geukensia demissa*. In *Encyclopedia of Life*. Retrieved March, 2016, from [http://www.eol.org/data\\_objects/33616650](http://www.eol.org/data_objects/33616650)

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., ... and Purcell, K. (2012). (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment* 10(6): 291–297. <http://dx.doi.org/10.1890/110236>

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J., & Collen, B. (2014). Defaunation in the Anthropocene. *Science*, 345(6195), 401-406

Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: MIT Press.

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., ... & Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges.

Ehrlich, P. R., Wilson, E. O. (1991). Biodiversity studies: science and policy. *Science*, 253(5021), 758-762.

Elo, S., and Kyngäs, H. (2008). The qualitative content analysis process. *Journal of advanced nursing*, 62(1), 107-115.

Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data—. *Ecological Informatics*, 11, 25-33.

EOL API. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://eol.org/api>

EOL Content Partners: Contribute Using Archives. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://eol.org/info/cp\\_archives](http://eol.org/info/cp_archives)

EOL Content Partners: Contribute Using EOL XML Transfer Schema. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://eol.org/info/create\\_xml](http://eol.org/info/create_xml)

EOL Content Partners: Contribute Using Spreadsheets. (n.d.). In *Encyclopedia of Life*. Retrieved March, Retrieved March, 2016, from <http://eol.org/info/337>

EOL History. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://www.eol.org/info/the\\_history\\_of\\_eol](http://www.eol.org/info/the_history_of_eol)

EOL Governance. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://www.eol.org/info/how\\_is\\_eol\\_managed](http://www.eol.org/info/how_is_eol_managed)

EOL Rubenstein Fellows Program. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://eol.org/info/fellows>

Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing?. *PLoS One*, 10(2), e0118053.

Fegraus, E. H., Andelman, S., Jones, M. B., & Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3), 158-168.

Fenichel, E. P., & Skelly, D. K. (2015). Why Should Data Be Free; Don't You Get What You Pay For?. *BioScience*, biv052.

Fernandez, J., Patrick, A., & Zuck, L. (2012). Ethical and Secure Data Sharing across Borders, in: Blyth, J., Dietrich, S., Camp, L.J. (Eds.), *Financial Cryptography and Data Security*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 136–140.

Flynn, F. J., & Chatman, J. (2002). " *What's the Norm Here?*": *Social Categorization as a Basis for Group Norm Development*. Division of Research, Harvard Business School.

Giant Panda. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://www.eol.org/pages/328070/overview>

Geschwind, D. H. (2001). Sharing gene expression data: an array of options. *Nature Reviews Neuroscience*, 2(6), 435-438.

Goring, S. J., Weathers, K. C., Dodds, W. K., Soranno, P. A., Sweet, L. C., Cheruvilil, K. S., ... & Utz, R. M. (2014). Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. *Frontiers in Ecology and the Environment*, 12(1), 39-47.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3), 156-162.

Harfoot, M., & Roberts, D. (2014). Taxonomy: Call for ecosystem modelling data. *Nature*, 505 (7482), 160. doi:10.1038/505160a

Hayes J. (2012). The data-sharing policy of the World Meteorological Organization: The case for international sharing of scientific data. Pages 29–31 in Mathae KB, Uhlir PF, eds. *Committee on the Case of International Sharing of Scientific Data: A Focus on Developing Countries*. National Academies Press.

He, Y. (2016, March). Star magnolia. In *iNaturalist*. Retrieved March, 2016, from <http://www.inaturalist.org/observations/2795649>

He, Y., Preece, J., Hammock, J., Butler, B., & Pauw, D. (2015, February). Understanding Data Providers in a Global Scientific Data Hub. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing* (pp. 215-218). ACM.

Hoegh-Guldberg, O., Mumby, P. J., Hooten, A. J., Steneck, R. S., Greenfield, P., Gomez, E., ... and Hatziolos, M. E. (2007). Coral reefs under rapid climate change and ocean acidification. *Science* 318(5857):1737-1742.

Hogg, M. A., & Terry, D. I. (2000). Social identity and self-categorization processes in organizational contexts. *Academy of management review*, 25(1), 121-140.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., ... & Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature*, 455(7209), 47-50.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conservation Letters*, 5(5), 399-406.

Ivezic Z. (2012). Data sharing in astronomy. Pages 41–45 in Mathae KB, Uhler PF, eds. Committee on the Case of International Sharing of Scientific Data: A Focus on Developing Countries. National Academies Press.

Jones, M. B., Berkley, C., Bojilova, J., & Schildhauer, M. (2001). Managing scientific metadata. *Internet Computing, IEEE*, 5(5), 59-68.

Jenkins, M. (2003). Prospects for biodiversity. *Science*, 302(5648), 1175-1177.

Jira : Project Management Software. (2016). Retrieved March, 2016, from <https://www.atlassian.com/software/jira>

Karasti, H., & Baker, K. S. (2004). Infrastructuring for the long-term: Ecological information management. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 1 (Vol. 1, pp. 10020c). Los Alamitos, CA: IEEE Computer Society.

Kaplan, B. (1991). Models of change and information systems research. In H.-E. Nissen, H. K.

Klein and R. Hirschheim (Eds.), *Information Systems Research: Contemporary Approaches and Emergent Traditions* (pp. 593–611). Amsterdam: Elsevier Science Publishers.

Kaye, J., Heeney, C., Hawkins, N., De Vries, J., & Boddington, P. (2009). Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331-335.

Kervin, K., Cook, R. B., & Michener, W. K. (2014, November). The Backstage Work of Data Sharing. In *Proceedings of the 18th International Conference on Supporting Group Work* (pp. 152-156). ACM.

Kim, S., Mankoff, J., & Paulos, E. (2013, February). Sensr: evaluating a flexible framework for authoring mobile data-collection tools for citizen science. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1453-1462). ACM.

Kim Y, Stanton JM (2012) Institutional and individual influences on scientists data sharing practices. *Journal of Computational Science Education* 3: 47–56.

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual review of information science and technology*, *45*(1), 247-294.

Kyngas, H., & Vanhanen, L. (1999). Content analysis. *Hoitotiede*, *11*(3-12).

Kratz, J. E., & Strasser, C. (2015). Researcher perspectives on publication and peer review of data. *PloS one*, *10*(2), e0117619.

Lauri, S., & Kyngäs, H. (2005). Developing Nursing Theories (Finnish: Hoitotieteen Teorian Kehittäminen). *Werner So derstrom, Dark Oy, Vantaa*.

Lee, C. P., Bietz, M. J., & Thayer, A. (2010, May). Research-driven stakeholders in cyberinfrastructure use and development. In *Collaborative Technologies and Systems (CTS), 2010 International Symposium on* (pp. 163-172). IEEE.

Lee, C. P., Dourish, P., & Mark, G. (2006, November). The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 483-492). ACM.

Learning and Education Working Group. (n.d.). In *Encyclopedia of Life*, Retrieved March, 2016, from [http://www.eol.org/info/learning\\_and\\_education\\_working\\_group](http://www.eol.org/info/learning_and_education_working_group)

Liittschwager, D. (2012). *A World in One Cubic Foot: Portraits of Biodiversity*. University of Chicago Press.

Linkert, M., Rueden, C. T., Allan, C., Burel, J. M., Moore, W., Patterson, A., ... & Swedlow, J. R. (2010). Metadata matters: access to image data in the real world. *The Journal of cell biology*, 189(5), 777-782.

Loarie, S. (2016a). About. In *iNaturalist*. Retrieved March, 2016, from <http://www.inaturalist.org/pages/about>

Loarie, S. (2016b). How can I use it. In *iNaturalist*. Retrieved March, 2016, from <http://www.inaturalist.org/pages/how+can+i+use+it>

Loarie, S. (2015). Supporters. In *iNaturalist*. Retrieved March, 2016, from <http://www.inaturalist.org/pages/partners>

Louv, R., & Fitzpatrick, J. W. (2012). Citizen science: public participation in environmental research. J. L. Dickinson, & R. Bonney (Eds.). Cornell University Press.

March, J. G., & Simon, H. A. (1958). Organizations.

Marco Rubio (speaker), Joe Pickens (Council Chair) (2008) Teacher Professional Development Programs in Florida. (Interim Project Report)  
<http://www.fldoe.org/core/fileparse.php/5636/urlt/0072410-professdevreport08.pdf>

McKeon, C., & Meyer, C. (2015). Award Abstract #1442731. Retrieved March, 2016, from [https://nsf.gov/awardsearch/showAward?AWD\\_ID=1442731](https://nsf.gov/awardsearch/showAward?AWD_ID=1442731)

Merriam, S. B. (2014). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, 7(1), 330-342.

Michener, W. K. (2006). Meta-information concepts for ecological data management. *Ecological informatics*, 1(1), 3-7.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, 27(2), 85-93

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, 10(6), 285-290.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. SAGE Publications, Incorporated. 3<sup>rd</sup> Edition.

Nadel, S. F. (2013). *The theory of social structure*. Routledge.

National Science Board. (2005). Long-lived digital data collections. Retrieved from <http://www.nsf.gov/pubs/2005/nsb0540/>

- National Science Foundation. (2010). NSF data sharing policy. Retrieved from [http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag\\_6.jsp#VID4](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4)
- National Science Foundation. (2015). NSF's public access plan: Today's Data, Tomorrow's Discoveries. <http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>
- Nielsen, M. A. (2012). *Reinventing Discovery: The New Era of Networked Science*. Princeton, NJ: Princeton University Press.
- Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment.
- Novacek, M. J. (2008). Engaging the public in biodiversity issues. *Proceedings of the National Academy of Sciences*, 105(Supplement 1), 11571-11578.
- Observations. (n.d.). In *iNaturalist*. Retrieved March, 2016, from [http://www.inaturalist.org/observations?taxon\\_name=](http://www.inaturalist.org/observations?taxon_name=)
- Ocasio, W. (1997). TOWARDS AN ATTENTION-BASED VIEW OF THE FIRM WILLIAM OCASIO. *Psychology*, 1, 403-404.
- Ocasio, W. (2011). Attention to attention. *Organization Science*, 22(5), 1286-1296

Parr, C. S., & Cummings, M. P. (2005). Data sharing in ecology and evolution. *Trends in Ecology and Evolution*, 20(7), 362-362.

Otegui J, Ariño AH, Encinas MA, Pando F. Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF). Raghava GPS, editor. PLoS One. 2013; 8: e55144.

doi: [10.1371/journal.pone.0055144](https://doi.org/10.1371/journal.pone.0055144) PMID: 23372828

Parr, C. S., Guralnick, R., Cellinese, N., & Page, R. D. (2012). Evolutionary informatics: unifying knowledge about the diversity of life. *Trends in ecology & evolution*, 27(2), 94-103.

Parr, C. S., Wilson, M. N., Leary, M. P., Schulz, K. S., Lans, M. K., Walley, M. L., ... & Corrigan Jr, M. R. J. (2014). The encyclopedia of life v2: providing global access to knowledge about life on Earth. *Biodiversity data journal*, (2).

Parr, C., Wilson, N., Schulz, K., Leary, P., Hammock, J., Rice, J., & Corrigan Jr, R. J. (2015). TraitBank: Practical semantics for organism attribute data. *Semantic Web*. Available at <http://www.semantic-web-journal.net/system/files/swj650.pdf>.

Piwovar, H. A., Becich, M. J., Bilofsky, H., & Crowley, R. S. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS medicine*, 5(9), e183.

Postmes, T., & Jetten, J. 2006b. Reconciling individuality and the group. In T. Postmes & J. Jetten (Eds.), *Individuality and the group: Advances in social identity*: 258-269.

London: Sage.

Pratt, M.G., & Rafaeli, A. 1997. Organizational dress as a symbol of multilayered social identities. *Academy of Management Journal*, 40(4): 862-898.

Preece, J., Sharp, H., & Rogers, Y. (2015). *Interaction Design-beyond humancomputer interaction*. John Wiley & Sons.

Project BudBurst. (2016). Annual Datasets. Retrieved March, 2016, from <http://budburst.org/datasets>

Rao, H., Monin, P., & Durand, R. (2003). Institutional Change in Toque Ville: Nouvelle Cuisine as an Identity Movement in French Gastronomy<sup>1</sup>. *American journal of sociology*, 108(4), 795-843.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, 331(6018).

Ribbed Mussel. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from <http://www.eol.org/pages/449853/overview>

Ribes, D., & Lee, C. P. (2010). Sociotechnical studies of cyberinfrastructure and e-research: current themes and future trajectories. *Computer Supported Cooperative Work (CSCW)*, 19(3-4), 231-244.

Rotman, D., et al. (2012). Supporting content curation communities: The case of the Encyclopedia of Life. *Journal of the American Society for Information Science and Technology*, 63(6), 1092-1107.

Rüegg, J., Gries, C., Bond-Lamberty, B., Bowen, G. J., Felzer, B. S., McIntyre, N. E., ... & Weathers, K. C. (2014). Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment*, 12(1), 24-30.

Ryan, Richard M. and Edward L. Deci (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68-78.

Schopf J.M, Bordenstein S, Leary P, Mangiafico P, Patterson DJ, Shipunov A, Shorthouse D (2008) Managing Biodiversity Knowledge in the Encyclopedia of Life. BNCOD 2008 Biodiversity Informatics Workshop, Cardiff University, 10th July 2008. 2 pp. URL: <http://biodiversity.cs.cf.ac.uk/bncod/SchopfEtAl.pdf>

Schmidt, B., Gemeinholzer, B., & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLoS ONE* 11(1): e0146695. doi:10.1371/journal.pone.0146695

Search TraitBank. (n.d.). In *Encyclopedia of Life*. Retrieved March, 2016, from [http://www.eol.org/data\\_search](http://www.eol.org/data_search)

Sherif, M. (1936). *The psychology of social norms*. New York: Harper & Row.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in ecology & evolution*, 24(9), 467-471.

Smith V.S., Rycroft S., Scott B., Baker E., Livermore L., Heaton A., Bouton K., Koureas D.N., Roberts D. (2012). Scratchpads 2.0: a virtual research environment infrastructure for biodiversity data. Accessed at <http://scratchpads.eu> on 2012-11-19

Soranno, P. A., Cheruvilil, K. S., Elliott, K. C., & Montgomery, G. M. (2015). It's Good to Share: Why Environmental Scientists' Ethics Are Out of Date. *BioScience*, 65(1), 69-73.

Species Pages Working Group. (n.d.). In *Encyclopedia of Life*, Retrieved March, 2016, from [http://www.eol.org/info/species\\_pages\\_working\\_group](http://www.eol.org/info/species_pages_working_group)

Star, S. L. (1991). The sociology of the invisible: The primacy of work in the writings of Anselm Strauss. In D. Maines (Ed.), *Social organization and social process: Essays in honor of Anselm Strauss* (pp. 265–283). Hawthorne, NY: Aldine de Gruyter, pp.265-283.

Star, S. L. (1999). The ethnography of infrastructure. *American behavioral scientist*, 43(3), 377-391.

Star, S. L., & Ruhleder, K. (1994, October 22–26). Steps towards an ecology of infrastructure: Complex problems in design and access for large-scale collaborative systems. *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW 94 – Transcending Boundaries)*, Chapel Hill, NC. ACM Press, New York, pp. 253–264.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7(1), 111-134.

Star, Susan; Griesemer, James (1989). "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39". *Social Studies of Science* 19 (3): 387–420.  
doi:10.1177/030631289019003001.

Steinhardt, S. B. (2016, May). Breaking Down While Building Up: Design and Decline in Emerging Infrastructures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2198-2208). ACM.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., ... & Fink, D. (2014). The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31-40.

Tajfel, H., & Turner, J. C. 1986. The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations*, 2nd ed.: 7-24. Chicago: Nelson-Hall.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... & Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PloS one*, *6*(6), e21101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., ... & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one*, *10*(8), e0134826.

The Taxon Page. (n.d.). In *Encyclopedia of Life*, Retrieved March, 2016, from [http://eol.org/info/taxon\\_page](http://eol.org/info/taxon_page)

Thornton, P. H. (2004). *Markets from culture: Institutional logics and organizational decisions in higher education publishing*. Stanford University Press.

Thornton, P. H., & Ocasio, W. (1999). Institutional logics and the historical contingency of power in organizations: Executive succession in the higher education publishing industry, 1958-1990. *American journal of Sociology*, 105(3), 801-843.

Thornton, P. H., & Ocasio, W. (2008). Institutional logics. *The Sage handbook of organizational institutionalism*, 840, 99-128.

Thornton, P. H., Ocasio, W., & Lounsbury, M. (2012). *The institutional logics perspective: A new approach to culture, structure, and process*. Oxford University Press.

Tyworth, M. (2014). Organizational identity and information systems: how organizational ICT reflect who an organization is. *European Journal of Information Systems*, 23(1), 69-83.

Ueda, K. (2016). API Reference. In *iNaturalist*. Retrieved March, 2016, from [https://www.inaturalist.org/pages/api reference](https://www.inaturalist.org/pages/api%20reference)

Van Horn, J. D., & Ball, C. A. (2008). Domain-specific data sharing in neuroscience: what do we have to learn from each other?. *Neuroinformatics*, 6(2), 117-121.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization science*, 16(4), 409-421.

What is EOL? - Encyclopedia of Life. (n.d.). Retrieved March, 2016, from

<http://www.eol.org/about>

Wiederhold, G. (1992). Mediators in the Architecture of Future Information Systems, *IEEE Computer*, Vol. 25, No. 3.

Weinstein, M. (2012). TAMS Analyzer for Macintosh OS X: The native Open source, Macintosh Qualitative Research Tool. Retrieved March, 2026.

Wiggins, A. (2012). Crowdsourcing Scientific Work: A Comparative Study of Technologies, Processes, and Outcomes in Citizen Science. *ProQuest LLC*.

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., Littauer, R., ... & Weltzin, J. (2013). Data management guide for public participation in scientific research. *DataOne Working Group*, 1-41.

Wiggins, A., & He, Y. (2016, February). Community-based Data Validation Practices in Citizen Science. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1548-1559). ACM.

Wiggins, A., & Preece, J. (2015). Award Abstract #1442668. Retrieved March, 2016, from [https://nsf.gov/awardsearch/showAward?AWD\\_ID=1442668](https://nsf.gov/awardsearch/showAward?AWD_ID=1442668)

Wilson, E. O. (2003). The encyclopedia of life. *Trends in Ecology & Evolution*, 18(2), 77-80.

Wolkovich, E. M., Regetz, J., & O'Connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, 18(7), 2102-2110.

Yin, R. K. (2013). *Case study research: Design and methods*. Sage publications.

Yang, Y., Tian, K., Hao, J., Pei, S., & Yang, Y. (2004). Biodiversity and biodiversity conservation in Yunnan, China. *Biodiversity & Conservation*, 13(4), 813-826.

Zimmerman, A. S. (2003). *Data sharing and secondary use of scientific data: experiences of ecologists* (Doctoral dissertation, The University of Michigan).

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology*, 58(4), 479-493.