2013-01-01

A Semantic Web-based Methodology For Describing Scientific Research Efforts

Aida Gandara
University of Texas at El Paso, aggandara@gmail.com

A Semantic Web-based Methodology For Describing Scientific Research Efforts

AIDA GANDARA

Department of Computer Science

APPROVED:

Natalia Villanueva-Rosales, Ph.D., Chair

Ann Quiroz Gates, Ph.D., Co-Chair

Christopher Kiekintveld, Ph.D.

Craig Tweedie, Ph.D.

Benjamin C. Flores, Ph.D.
Dean of the Graduate School

To Waldo, Isa, Saaco and Nano,

... we're not inclined to resign to maturity!

A Semantic Web-based Methodology For Describing Scientific Research Efforts

by

AIDA GANDARA, B.S, M.S

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science

THE UNIVERSITY OF TEXAS AT EL PASO

August 2013

Acknowledgements

I received support from many people and organizations during my doctoral studies. These people deserve more than this simple acknowledgement. I hope that I can continue to show them the value of what they have done through my future endeavors.

First, my family. Edgar Gándara, my husband, had to wear many 'hats' so I could work, especially for my final hiatus from normal life. In addition, Edgar was supportive during the ups and downs throughout my studies and he never doubted my capabilities. My three children, Marisa, Isaac and Edgar Iván, you three inspire me; each so unique, so caring and so funny - you kept me going! My mother, Sylvia Portillo, who not only sent delicious weekly meals and gave rides to my kids, she shared her contagious love of books and learning with me throughout my life. My wonderful siblings and friends, Sylvia, Albert, Louie and Diana, who assure that all our holidays and reunions are filled with laughter, good food, and more laughter; I needed all of those moments to lighten the load. My dad, Alberto José Gutiérrez, who was not able to make it through this journey with me, taught me to stay strong with my faith especially when life is hard. I feel like we have all earned this PhD.

There are several people who have come and gone at the Cyber-ShARE Research Center, my research home. Working with them **all** has made these years unique and more enjoyable. My advisor, Dr. Natalia Villanueva-Rosales, entered at the end of my research and shared her thoughts, inspired my thinking and helped me mature as a researcher. She naturally functions as a catalyst and her support for students and innovative research has already started to change the dynamics within the Cyber-ShARE Knowledge Group. Dr. Ann Quiroz Gates has been an inspiration to me all these years, first, in supporting my decision to pursue a PhD and then with her continued support through the different decisions and challenges that I encountered. Ann's impact on the lives of many people at Cyber-ShARE, UTEP and around the country is not really measurable and there will continue to

be successes attributed to her efforts in broadening participation and computing research. Completing my doctoral degree is just one of them. Dr. Deana Pennington guided my research path by so openly sharing her experiences, vision and ideas. She influenced my perspective on innovation and collaboration between scientists and computer scientists. Deana is continuously engaging students at the Center in innovative and collaborative research and many of us continue to benefit from the opportunities and encouragement she provides. Patricia Esparza helped with the initial implementation of this research and has been a good friend since. Mary Contreras was supportive with handling paperwork for travel, reimbursements and appointments so I could focus on research. Aline Jaimes, Dr. Anibal Sosa and Lennox Thompson shared the details of their work with me. Dr. Leonardo Salayandía and Alla Dove proofread some sections of my dissertation. Thank you all.

The additional members of my dissertation committee, Dr. Craig Tweedie and Dr. Christopher Kiekintveld, have been supportive and insightful. Their time is valuable and I appreciate that they sacrificed some of it for this research.

I have a long history with the faculty in the UTEP Computer Science Department. Through them I began my teaching and interest in research. They challenged me academically and welcomed me as a professor and student for many years. I look forward to their continuing success.

Several organizations supported me financially. Cyber-ShARE (NSF CREST Grant No. HRD-0734825, HRD-1242122), Virtual Learning Commons (NSF CI-Team Grant OCI-1135525), Alliances for Graduate Education and the Professoriate (AGEP) and the Patricia and Jonathan Rogers Scholarship in Graduate Engineering funded my education. Completing this degree was possible because of their investment in me. Through the mentoring of Dr. Terence Critchlow and Dr. George Chin from the Pacific Northwest National Laboratory‘s (PNNL) SciDAC Group, I participated in an enlightening experience in professionalism and scientific innovation. Similarly, Hilmar Lapp from the National Evolutionary Synthesis Center (NESCent), who mentored me on an internship with DataONE, shared his sharp thinking and innovative ideas for Semantic-Web based information management.

Finally, to my many friends who are always there for me to share a coffee, laugh with, enjoy games, carpool and discuss the ups and downs of life. I both needed and enjoyed those times during my studies.

NOTE: This dissertation was submitted to my Supervising Committee on June 14, 2013

Abstract

Scientists produce research resources that are useful to future research and innovative efforts. In a typical scientific scenario, the results created by a collaborative team often include numerous artifacts, observations and relationships relevant to research findings, such as programs that generate data, parameters that impact outputs, workflows that describe processes, and publications, posters and presentations that explain results and findings. Scientists have options in what results to share and how to share them, however, there is no systematic approach to documenting scientific research and sharing it on the Web.

The goal of this research is to define a systematic approach for describing the resources associated with a scientific research effort such that results and related resources become more accessible and understandable to machines over the Semantic Web. This research defines a methodology, called **C**ollect-**A**nnotate-**R**efine-**P**ublish Methodology (CARP), that uniformly structures and links heterogeneous information that is distributed over the Web as scientific collections. Scientific collections are structured descriptions about scientific research that make scientific results accessible based on context and queryable by machines. Initial findings confirm that documenting scientific research on the Web is not a common practice and that tools that implement CARP can guide the process and facilitate documenting scientific research on the Linked Data dataspace. As a result, machines can help understand resources and their relationships in scientific collections, therefore, CARP facilitates reuse of scientific results.

Table of Contents

Page

Acknowledgements v

Abstract viii

Table of Contents ix

List of Tables xii

List of Figures xiii

Chapter

1 Introduction 1

1.1 Motivation 1

1.2 Goal and Objectives 5

1.3 Intellectual Merit and Broader Impact 6

2 Background 8

2.1 The Semantic Web 8

2.2 Ontologies 11

2.3 Semantic Web Servers 15

2.4 Linked Data 18

2.5 An Illustration of the Linked Data Principles 19

2.6 Global Linked Data Dataspace 22

3 The Approach 25

3.1 Guiding Principles 25

3.1.1 Principle 1: Relate resources to a collection 26

3.1.2 Principle 2: Reuse existing Web content 27

3.1.3 Principle 3: Employ defaults 28

3.1.4 Principle 4: Capture explicit relationships 29

3.1.5 Principle 5: Enable machines Web access 30

3.2 Significance of Guiding Principles . . . 31
3.2.1 A Structured Information Base . . . 31
3.2.2 Issues . . . 33
3.3 The CARP Methodology . . . 38
3.3.1 Overview . . . 38
3.3.2 CARP Phases . . . 39
4 Implementation of CARP Methodology . . . 50
4.1 CARP Prototype . . . 50
4.2 Related Approaches to Support CARP . . . 60
4.3 Summary . . . 71
5 Validation and Verification . . . 73
5.1 Case Studies . . . 73
5.1.1 Eddy Covariance Cyber-infrastructure (ECC) Project . . . 77
5.1.2 Receiver Function Modeling (RFM) Project . . . 88
5.1.3 Constraint Optimization (CO) Project . . . 97
5.1.4 Summary . . . 104
5.2 A Survey on Scientific Collections . . . 106
6 Discussion . . . 110
6.1 Research stages . . . 110
6.2 Lack of evidence . . . 111
6.3 Attribution and licensing policy . . . 111
6.4 RDFization or Self-description . . . 112
6.5 Publication tools and self-descriptions . . . 113
6.6 An evolving Semantic Web . . . 113
6.7 Semantic Web investment . . . 115
7 Summary . . . 116
7.1 Overview of Work . . . 116
7.2 Future Work . . . 117

7.3 Concluding Remarks 119
References 122
Glossary 135
Acronyms 137
Appendix
A The Initial Methodology 139
B An Initial User Survey 142
Curriculum Vitae 147

List of Tables

4.1 The table considers how each approach collects, annotates, relates and publishes information on the Web or Semantic Web. 70

5.1 The two tables summarize the state of resources at the beginning of documenting the ECC project case study. The top table shows the different resources, where they were located, their attribution and self-description. The bottom table shows the collaborator information calculated from the resources, including self-description, the collaborations found in some online document and participation loss. 81

5.2 ECC Resource Types 83

5.3 ECC Properties 84

5.4 Two tables summarizing the state of resources at the beginning of documenting the RFM project case study. The top table shows the different resources, where they were located, their attribution and self-description. The bottom table shows the collaborator information calculated from the resources, including self-description, the collaborations found in some online document and participation loss. 91

5.5 RFM Resource Types 93

5.6 RFM Properties 94

5.7 Tables showing resources and collaborators of the CO project 99

5.8 CO Resource Types 101

5.9 CO Properties 102

List of Figures

2.1 A publication about joint inversion with a URI to identify it as a Web resource 11
2.2 A publication with a URI to identify it as a Web resource. The values below the URI are properties that describe the resource 11
2.3 Two classes found in the BibTex Ontology, Article and Entry. The arrow connecting the two classes signifies that Article is a subClass of Entry. The blue box shows that Article inherits both class and properties from the Entry class 13
2.4 The joint inversion publication semantically described using the BibTex Ontology. The resource is described as a BibTex Article instance with Bibtex **has keyword**, **has year** and **has publisher** properties. The format property is a Dublin-Core term that can be applied to any class. 14
2.5 A diagram representing scientific resources. Some comply with Linked Data principles others are not even Web accessible. 21
2.6 A representation of the Lined Open Data Cloud from 2011. The nodes represent organizations that have shared data on the Linked Data Cloud using Linked Data principles. 23

3.1 The resources in a scientific collection reuse existing Web information, add additional information to describe resources, and add relationships between resources. 27
3.2 CARP principles assure that the creation of collections of information on the Web reuse, preserve and expose structured information so it is compatible with the Linked Data dataspace. 35

3.3 The four phases of CARP: Collect, Annotate, Refine and Publish, each shown in an oval with arrows that designate the flow of information to and from a scientific collection. The scientific collection is in the middle. 40
3.4 A graph representing a scientific collection that reflects Figure 3.1. In the ovals are the named individuals and the gray boxes represent roles. The arrows signify the relationship between objects or values. Not all properties or individuals are shown, for readability. 49

4.1 Sample fields of a Drupal node to an RDF graph 53
4.2 The CI-Server multiple triplestore architecture. Each scientific collection is held in a triplestore that supplements the Drupal database. Scientific collections can be accessed and queried individually through a SPARQL-endpoint. 54
4.3 The core CARP vocabulary defines classes and properties used by the CI-Server implementation of CARP. CARP concepts inherit from other ontologies; respective namespaces are listed in the Namespaces box. The arrows between each class represent relationships, identified by the labels. 56
4.4 The CI-Server Framework on the Drupal CMS. Nodes and web-accessible URIs are RDFized and stored in an ARC2 triplestore, representing a scientific collection. Comments, annotations and other refinements are also added as triples into the collection. A SPARQL-endpoint and self-describing URI are exposed on the Semantic Web. 59
4.5 CARP reuses information on the Web making scientific collections accessible as Linked Data. 72

5.1 Summary of meetings held with the different case studies. 77
5.2 Quality Control and Gap Filling resources and relationships documented in the research effort 87

5.3 A graph representing the results from querying the ECC scientific collection about **all** contributors. 88
5.4 A graph representing the results from querying the ECC scientific collection about **only** poster contributors. 88
5.5 Image showing all receiver function locations 96
5.6 Map generated from self-describing a station table 96
5.7 PDIP output for the Archean model, including comments. 105
5.8 Responses about accessibility of resources 108
5.9 Responses about understanding of resources 108
5.10 Responses about attribution of resources 109
5.11 Responses about notes and discussions about resources 109

6.1 Self-describing poster template that uses default vocabulary and guides the structured documentation of research resources. 114

A.1 The initial methodology 141

B.1 Results for the initial survey on accessibility 144
B.2 Results for the initial survey on understanding of resources 145
B.3 Results for the initial survey on attribution of resources 146

Chapter 1

Introduction

Scientists produce research resources, e.g., publications, posters, datasets, images, etc., that are useful to future research and innovative efforts. With the many options available for sharing scientific results, finding and understanding research related resources that characterize the results of a research effort is a challenge. Solutions are needed to explore and exploit the deluge of data from the scientific community, considering techniques that enable scientific experts, not necessarily computer or data experts, to expose related information and its meaning (hey, 2005; Fox and Hendler, 2009). The Semantic Web, through the use of uniform data representation and protocols to access World Wide Web (Web) resources can mitigate issues with accessing and understanding related information distributed and heterogeneously available over the Web. This research defines a methodology for sharing scientific results as a collection over the Semantic Web, consistently creating a more comprehensive and uniformly structured description of the results of a scientific research effort. This research argues that the methodology facilitates how resources are accessed and how machines can help in understanding them.

1.1 Motivation

Scientists have varied options in what results to share and how to share them. Some of these approaches (Pearce et al., 2010; Kraker et al., 2011; De Roure et al., 2007; van Sompel and Lagoze, 2009) fall within the scope of Science 2.0, an enhanced version of scientific research that steps out of former rigid laboratory research methods to capture observations, interactions and data in real world environments using socio-technical systems (Shneider-

man, 2008). Aside from publications that document scientific research, current approaches to sharing scientific results are primarily focused on sharing data, e.g., raw or curated data in digital format, and, in some cases, accompanying metadata. Some data are available from Web servers setup by researchers to describe and share research resources, e.g., by enabling FTP[1] downloads. Scientists also share data over the Web at data portals or library information managers, as part of a data management plan or to contribute to the holdings of a specific research community (IRIS, 2012; The Knowledge Network for Biocomplexity, 2012; ORNL, 2012).

In a typical scientific scenario, the results created by a collaborative team often include numerous artifacts, observations and relationships relevant to research findings, such as programs that generate data, parameters that impact outputs, workflows that describe processes, and publications, posters and presentations that explain results and findings. Workflows and related data can be published at the myExperiment portal, a collaborative scientific workflow environment created to share, discuss and rate workflows that capture the algorithmic process for conducting scientific experiments (De Roure et al., 2007). Programs can be uploaded to source code repositories suh as Github (Dabbish et al., 2012), enabling other scientists to reuse or re-purposec the code. Some research results are not publicly available, for a variety of reasons, prompting investigations and unified efforts into facilitating the general availability of such resources on the Web (Reichman et al., 2011).

One disadvantage to the various mechanisms for sharing research results on the Web is the lack of interoperability for users or software agents to access and understand them. Each data portal, for example, uses different metadata to capture data and provides different techniques to access managed information, i.e., by providing tools or interfaces that enable access to data and metadata through online searches to HTML or XML formatted Webpages. The Alaska Data Integration Working Group (ADIwg) (ADIwg, 2013) is a unified effort of various organizations formed to examine and address the technical barriers to integrating and sharing data across various data portals that capture data related to

[1]FTP (file transfer protocol) is a protocol for downloading and uploading files.

the Alaska region. As a result of these efforts, ADIwg has identified standard metadata and Web services to uniformly capture and exchange published data. Similarly, the Data Observation Network for Earth (DataONE) (Strasser et al., 2011) is a unified effort identifying a multiple hierarchy network for exchanging data across participating data providers. The DataONE participating organizations share tools and techniques created to provide data, use common APIs for searching and exchanging data, and continually enhance their integration efforts across the organization (Strasser et al., 2011). Despite unified efforts such as the two mentioned above, the Web consists of a melange of Websites, metadata standards, APIs and services that exacerbate and already tedious process of understanding scientific research. Users and software agents interested in understanding the resources available across the Web must access each data provider and understand the nuances of different interfaces in order to view, compare, integrate and download resources.

Another disadvantage to distributed approach of sharing resources across the Web is that the implicit relationships inherent in research results, such as those between data, people, publications and programs that exist during scientific research are usually not expressed on the Web. Some Web data providers expose links to related information, for example the USArray[2] shares hyperlinks to information related to certain topics, e.g., station lists, station maps and station reports. Hyperlinks, however, are to Web pages requiring further analysis to identify how stations relate to station maps. Dryad.org[3] shares hyperlinks between data and publication Digital Object Identifiers (DOIs). The hyperlinks are relationships between two resources from different data sources. Since hyperlinks are unnamed, they lack a meaningful relationship. One requirement for publishing a dataset on Dryad.org is that the corresponding publication be peer reviewed. Researchers may produce many datasets that are accessible on the Web but not related to a peer reviewed publication, and these are not qualified for publication at the Dryad.org Website.

Collaborative scientific efforts, such as those carried out at the Cyber-ShARE Research

[2]USArray is a data provider of seismic data; http://www.usarray.org/.

[3]Dryad.org is a data provider of ecological data that has a related peer reviewed publication; http://datadryad.org/.

Center of Excellence[4] (Cyber-ShARE), would benefit from increasing the understanding of details about these collaborations beyond the written reports to funding agencies and publications. Currently for the Center, sharing of results occur through publications, sparsely through data portals and ongoing through reports, presentations and meetings. Sharing more details about scientific research would expose details about individual research efforts funded by the Center. Scientists face a deluge of data beyond what can be managed by tools today due to the complexity in data capture, complexity in analysis, and the size, distribution and heterogeneity of data. To make scientific research findings more understandable and reusable, there is a need to develop technologies to facilitate the sharing of information about the research process as well (Gray et al., 2005). Technologies to help on this front must lower the barrier to accessing and understanding distributed and distinctly available holdings of scientific research results (Fox and Hendler, 2009). Moreover, for Cyber-ShARE research collaborations, approaches for sharing scientific research over the Web should enable scientists to share a comprehensive collection of research related results such that: 1) resources can exist at distributed locations on the Web yet still be directly accessible from research documentation, 2) metadata about research related resources can be queried or analyzed despite the actual location or format of resources, 3) documentation of resources is meaningfully structured such that it can be integrated and compared to other research efforts.

The Semantic Web was introduced to facilitate processing over the growing amount of Web information while reducing the reliance on user intervention (Berners-Lee et al., 2001) by making Web content machine understandable, despite it being distributed or heterogeneous in structure. Using Semantic Web techniques, Web resources are described through well-defined structures enabling people to create tools that reliably process them (Berners-Lee et al., 2001). Linked data, a Semantic Web-based approach, identifies principles for sharing data over the Web where structured data is uniquely accessible to humans

[4]The Cyber-ShARE Center of Excellence is supported by National Science Foundation grant number HRD-0734825.

or machines and relationships are typed between different data sources (Bizer et al., 2009a). Using Linked Data approaches can alleviate challenges for researchers to describe research results currently shared over the Web, separate from the nuances of distributed and heterogeneous data holdings. However, previous work with documenting Cyber-ShARE research efforts conclude that there is no systematic approach to documenting scientific research or sharing it on the Web or Semantic Web (Gándara, 2012; ci1, 2012; Gándara et al., 2011b).

1.2 Goal and Objectives

The **goal of this research** is to define a systematic approach for describing the resources associated with a scientific research effort such that results and related resources become more accessible and understandable to machines over the Semantic Web. Toward this goal, the *objectives* of this research are to:

1. *design* a methodology for describing scientific research as a scientific collection of Web resources such that the Web resources are semantically described using existing Web content, relationships between Web resources have been explicitly identified and the collection is searchable over the Semantic Web. The activities involved with this objective are as follows:

    - identify steps to semantically describe Web resources within a research effort; and
    - identify the steps to capture relationships between resources within a scientific collection; and
    - identify the steps to expose the scientific collection as searchable over the Semantic Web.

2. *validate* the effectiveness of the methodology to create scientific collections of scientific Web resources, explicitly identify relationships between Web resources, and expose

the scientific collection as searchable on the Semantic Web. The activities involved with this objective are as follows:

- create a prototype system that implements this methodology; and
- document three case studies from research conducted at Cyber-ShARE using the prototype.
- conduct a survey on accessibility, understanding, and attribution of scientific collections with Cyber-ShARE researchers.

3. *verify* that scientific collections are machine understandable. The activities involved with this objective are as follows:

   - identify research questions relevant to each one of the three case studies; and
   - translate research questions into machine understandable queries; and
   - apply machine understandable queries to scientific collections and verify the results.

1.3 Intellectual Merit and Broader Impact

The **intellectual merit** of the research is that it defines a methodology for creating scientific collections over the Semantic Web and for increasing accessibility of scientific research results and relevant information to support reuse. The methodology allows heterogeneous information that is distributed over the Web to be uniformly structured and linked. The descriptions of machine-processible information facilitates discovery and integration with other Semantic Web information, increasing understanding of scientific research results for collaboration and reuse. The research discusses the interaction of existing Semantic Web tools and techniques that are used together to create scientific collections. By using generic Semantic Web techniques, collections can be used to describe almost any topic, not just

scientific resources, facilitating access and understandability of many topics from information currently shared on the Web.

The remainder of this manuscript covers the following: Chapter 2 presents background information to understand current challenges in sharing research results over the Web; Chapter 3 introduces the guiding principles for defining a methodology that supports discovery, reuse and integration of Web resources within scientific collections; Chapter 4 introduces the methodology called the **C**ollect-**A**nnotate-**R**efine-**P**ublish Methodology (CARP) that defines the process for creating scientific collections; Chapter 5 describes a prototype system that implements the methodology; Chapter 6 presents the validation and verification of the methodology; Chapter 7 evaluates the guiding principles and the methodology; and Chapter 8 discusses findings and lessons learned from documenting scientific research; and Chapter 9 presents conclusions for the work, research outcomes and discusses future work. Additionally, Appendix A and B present supplementary information; Appendix A discusses an initial methodology that precluded CARP; and Appendix B presents a small survey that shares an initial understanding into the types of resources managed by researchers at Cyber-ShARE as well as their opinions on the importance of sharing research resources.

Chapter 2

Background

The research presented leverages tools and techniques of the Semantic Web to both describe and share information about scientific research. This chapter first describes the Semantic Web and how it functions to uniformly describe Web resources that are heterogeneous and the role that ontologies and Semantic Web servers play in how the Semantic Web works. Then, Linked Data and its role in building a global dataspace is discussed.
Notice: for clarity, in the remainder of the manuscript, the **bold** emphasis is used to depict ontology terms.

2.1 The Semantic Web

The Web has become a common mechanism for sharing information. To facilitate working with the proliferation of Web information, machines have been enhanced to work in ways once handled by humans. For example, accessing a website and extracting needed content once required a person to logon to a Website, if it was secured with a password, search for the content and download it or open it up, all manually. Web information extraction tools (Chang et al., 2006) extract content from HTML Web pages in order to process the information, similar to how a human might identify information visually from a Web page. Extraction tools can follow links and continue extracting information, mimicking the human ability to surf the Web. Web search tools are implemented with algorithms to find the 'best match'. The PageRank algorithm, introduced in the google Web search engine (Brin and Page, 1998), might find information relevant to a user's query but the user must still determine the appropriateness of returned information. In addition, despite finding a

URL with search related content, tools to view the content of a URL may not be available to the searching user. Likewise, a user can search for URLs and may never find the content of interest.

The Semantic Web was introduced specifically for alleviating such search problems and facilitating machine processing of Web information. The Semantic Web was introduced to make Web content machine processible, despite it being distributed or heterogeneous in structure, so as to reduce the reliance on user intervention (Berners-Lee et al., 2001; W3C, 2012). In 2001, the vision of the Semantic Web was introduced, illustrating an interactive and meaningful experience where users are able to seamlessly access information, relying on their computer's ability to understand and process the variety of information available on the Web (Berners-Lee et al., 2001). The Semantic Web is described as an extension of the existing Web where ontologies provide meaningful descriptions of Web resources. The descriptions are accessible to software agents that are able to understand their meaning, rather than just display information for human processing. The Semantic Web community is still pursuing this vision (W3C, 2012; Shadbolt et al., 2006; Fox and Hendler, 2009) with increasing participation in the Semantic Web community.

The Semantic Web exposes meaning over the Hypertext Transfer Protocol (HTTP) protocol(Fielding et al., 1999), through URIs, unique identifiers that identify Web resources. There are several protocols that can be referenced via a URI, however, the Semantic Web functions over HTTP. URIs can reference anything, web-accessible or not, e.g., people, digital documents or emotions. URLs are URIs that reference documents located on the Web, e.g., a Web page. Not all URIs are documents, some reference data that may be located on a fileserver, database or calculated dynamically. A Web server is a software program that dereferences URIs to Hypertext Markup Language (HTML) or some other format, e.g., 'pdf' or 'jpg', based on a request of a Web client tool or software agent, e.g., a Web browser (Ding et al., 2004; SemanticWeb.org, 2013). The process of resolving a URI to a Web resource is called dereferencing the URI and the process of returning a URI request to a specific format is called content negotiation (Berners-Lee et al., 2005). Figure 2.1

shows a scholarly publication about joint inversion, a process used by scientists to jointly invert two types of data to obtain a single model of a structure, e.g., of the earth, the human body or other physical systems (Haber and Oldenburg, 1997). The publication's URI is dereferencible by a Web server. When a client tool accesses the URI the document is returned, i.e., the Web server 'serves' the Web content. Since there can be several representations of the same URI, the client and server must negotiate the content based on the client's request. The Semantic Web is described as a 'layer' over the Web because it uses information on the Web as the basis for semantically describing the Web. When client tools request a URI they can request the semantic description as the content type the a Web server should return.

Resources are semantically described through the use of structured languages, called ontologies, that provide meaning through classification and properties (Baader et al., 2010; Brickley and Guha, 2004). As seen in Figure 2.2, the scholarly publication can be enhanced with a meaningful description that identifies it to be of **Type**=article, **Keyword**=joint inversion, **Format**=acrobat pdf, **Year**=1997 and **Publication**=Inverse Problems, a journal publisher[1]. When one of the formats returned for a URI through content negotiation is a semantic description then the URI is a self-describing URI. A client tool can request a semantic description of a resource in order to obtain more understanding about what the resource is before trying to open or process the resource's content. In the case of the publication found in Figure 2.2, the description would be the properties specified below the publication's URL, such as **Type** and **Keyword**, and so on. On the Semantic Web, client tools aim to process the meaning of a resource.

A search for information on the Semantic Web can return more relevant results than current web-based techniques because self-describing properties are leveraged in the query. If a browser is used to search the Web for scholarly publications about joint inversion, the result may be several million URLs, some papers about joint inversion while others

[1]Inverse Problems publishes mathematical and experimental papers on inverse problems. http://iopscience.iop.org

URL = http://pubs/p01

Figure 2.1: A publication about joint inversion with a URI to identify it as a Web resource

URL = http://pubs/p01

Type: article
Keyword: joint inversion
Format: acrobat pdf
Year: 1997
Publication: Inverse Problems

Figure 2.2: A publication with a URI to identify it as a Web resource. The values below the URI are properties that describe the resource

could be Web pages, authors or other resources that are somehow related to the terms 'joint' and 'inversion'. If a semantically-aware browser is used to search the Semantic Web for scholarly publications about joint inversion, the description of Web content can be searched for resources having specific properties, e.g., **Type=article** and **Keyword=joint inversion**. The query result can return many resources all with the requested properties. Further processing can limit the result resources to formats the client tool can process, such as **Format=pdf**.

2.2 Ontologies

Ontologies are leveraged on the Semantic Web to formally describe concepts and relationships about data, together referred to as the conceptualizations of the ontology (Gruber, 1995). Typically, ontologies are created with ontology development tools or ontology editors (Noy et al., 2001; Kalyanpur et al., 2006; Day-Richter et al., 2007; Bernstein and Kaufmann, 2006) to define the terms (vocabulary) of the concepts described by the ontology using the constructs of an ontology language (Gómez-Pérez, 1999), e.g., the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2003) or Resource Description Frame-

work Schema (RDF-S) (Brickley and Guha, 2004). The more common terms to identify in an ontology are *classes* that describe concepts like **car** or **book**, *subclasses*, *data properties* that describe a relationship between a class and basic data values like strings and numbers, *object properties* that describe relationships between classes and other classes, and *subproperties*. Some ontology editors also handle details in ontology languages to fulfill requirements on the terms described by the ontology, for example, property characteristics such as inverse or property restrictions such as cardinality. The two parts of an ontology are the description of the terms and, optionally, the individuals or instantiations of classes.

There are numerous ontologies currently defined, describing physical objects like people and research communities (Brickley and Miller, 2010; Sure et al., 2005), digital objects like publications (Knouf, 2004) and abstract objects like emotions (López et al., 2008). Figure 2.3 shows a class **Article** as a subclass of **Entry** class, both defined in the BibTex Ontology[2]. BibTex is a reference management tool that structures lists of references using types and properties for different publication types, however, it is not structured with a known Semantic Web description language, thus, the BibTex Ontology introduces the structure of BibTex using OWL classes and properties. Details about the **Article** class, shown in the blue box, describe the properties that are inherited from the **Entry** class, i.e., **has year** and **has title**. In addition, not shown, are additional properties from the ontology such as **has publisher** that represents where the resource is published and **has keyword** to identify keywords that characterize the resource. Self-describing URIs are instantiations of ontologies. Thus, the joint inversion publication can be described by the BibTex Ontology as shown in Figure 2.4, by identifying the resource as a BibTex Article and setting relevant properties. The diversity in types and amount of ontologies currently available has prompted research in ontology libraries and ontology search tools (d'Aquin and Noy, 2012; Ding et al., 2004) to make ontologies easier to find. There is no requirement to use a specific ontology for describing content on the Semantic Web, however, using terms from the same ontology can facilitate searches for similar resources. In addition, using the

[2]The BibTex ontology can be found at: http://zeitkunst.org/bibtex/0.2/bibtex.owl

Figure 2.3: Two classes found in the BibTex Ontology, Article and Entry. The arrow connecting the two classes signifies that Article is a subClass of Entry. The blue box shows that Article inherits both class and properties from the Entry class

same terms as other Web resources can facilitate information integration of distributed information (Shadbolt et al., 2006; Noy, 2004).

Ontologies can be created about any topic and published on the Web. New ontologies are normally created to add terms or extend existing terms, through subclasses and sub-properties, and to create new individuals. To share an ontology on the Web, the language constructs are serialized to a file that can be assigned a URL. An ontology can reuse terms from another ontology by importing the ontology. With the OWL and RDF-S ontology languages, ontologies are serialized to an Extensible Markup Language (XML)-based syntax called Resource Description Framework (RDF) (Klyne and Carroll, 2004) that works with ontology constructs as triples, a tuple-based approach to uniformly work with structured information. RDF is an abstract data model, often referred to as a language (Klyne and Carroll, 2004), for describing semantic information as object-attribute-value triples that are syntactically represented using XML. RDF is domain independent so users must leverage other languages such as RDF Schema and OWL to express domain specific concepts such as classes and properties. These more expressive languages build upon RDF, hence, using

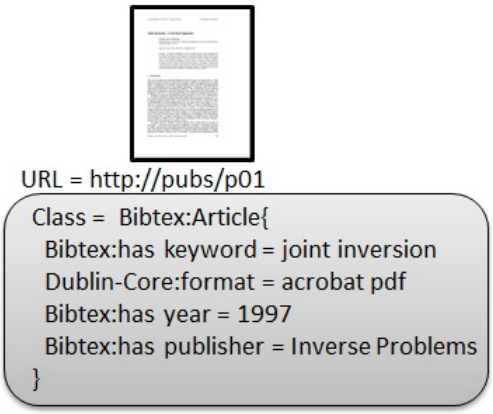

Figure 2.4: The joint inversion publication semantically described using the BibTex Ontology. The resource is described as a BibTex Article instance with Bibtex **has keyword**, **has year** and **has publisher** properties. The format property is a Dublin-Core term that can be applied to any class.

RDF triples and XML to represent and exchange information supports changes in the vocabularies of more expressive languages without requiring all data consumers, i.e., Semantic Web clients or software agents, to be changed (Antoniou and van Harmelen, 2008).

When an ontology does not have the necessary classes or properties to describe a conceptualization, terms can be used from other ontologies or new terms can be added to an existing ontology. Notice that the semantic description of the article in Figure 2.4 only includes three BibTex fields; there is a fourth labeled Dublin-Core. The BibTex ontology only defined three of the properties that were needed for the joint inversion publication's semantic description. There are two choices for fixing such an issue: a new ontology can be created and the **Article** class extended to allow for a **format** property. Alternatively, terms from another ontology can be used, as long as ontology constraints are followed. For example, if an ontology species that a property can only be applied to a resource of class **book**, then it should not be applied to a resource of class **Article**. The **format** property can be applied to any resource class so it is used to describe the joint inversion publication.

There are challenges that limit Semantic Web deployment (Heath, 2009; Cardoso, 2007). For example, choosing how to create and use ontologies has proven to be a challenge for

the Semantic Web community, in particular, because the design of an ontology can be delayed when a group can not conform to a single vocabulary. Similarly, the Semantic Web works to address deployment issues, e.g., research in pay-as-you-go ontology integration is focused on mitigating some ontology design delays by encouraging the use of common vocabularies and proprietary terms and deferring schema decisions that assure integration other Semantic Web vocabularies until necessary (Franklin et al., 2005; Das Sarma et al., 2008).

2.3 Semantic Web Servers

Web servers that manage and share semantic information described with ontologies are referred to as Semantic Web portals and Semantic content management systems, similar to their Web counterparts. For example, a Semantic Web portal functions similar to a Web portal by providing tools and services from a single Web location (Davies et al., 2003). Semantic Web servers play a crucial role in the availability of semantic information on the Web and as such, must be instrumented to collect and publish semantic information if the Semantic Web vision is to be realized. Semantic Web servers import ontologies to provide the structure of information managed by the server (Reid and Edwards, 2009; Gándara et al., 2011b). Changes to an ontology occur at the source of the ontology, i.e., by an editor that changes the original ontology and exports it to be shared on the Web. Loading ontologies can occur seamlessly, where services or other mechanisms are employed to update the ontology automatically (Davies et al., 2003) or loading an ontology can be a manual step where a system administrator or user loads the ontology and aligns it with details in the system (Corlosquet et al., 2009). Semantic Web servers should control the definition of an ontology created local to the server, within the server's namespace. A namespace is a container for a set of identifiers (names) that provides disambiguation of homonym identifiers that reside in different servers. On the Web, a namespace is a naming technique provided relative to a Web server's name. For example, http://cybershare.utep.edu

is the Cyber-ShARE namespace and all Web resources that are introduced by Cyber-ShARE would begin with the namespace and should be accessible from the Cyber-ShARE server. In addition, dereferencing and content negotiation of namespace resources should be managed by the Cyber-ShARE server, making it the authority of all resources served by the http://cybershare.utep.edu namespace. Using these conventions simplifies identifying where a resource should be found or modified.

The dependence on external sources to define the structure of information, i.e., provide ontologies, used in a Web server has some positive and some negative qualities. One positive quality is that the Web server is importing formalized descriptions of information that can cater to a specific domain and community of users. If the Web server is managing information about events and the people involved in those events, it can load and work with a description specifically defined by the community that needs it, not a predefined structure defined by the Web server. Also, ontologies can be loaded dynamically, enabling the description of information to change as needed by the ontology community, albeit, through the two step process of using an ontology editor external to the Web server and requiring the knowledge or intuition to know when to load updated ontologies. Finally, Semantic Web servers that enable users to control the structure of managed content, can also enable users to control the structure of information that is shared on the Semantic Web because a Web server can use those ontologies for self-describing URIs. On the downside, if ontologies must exist for a user or a community of users to take advantage of Semantic Web servers, then users must first identify the ontology that suits their needs. Identifying and reusing ontologies is a challenge to users who have little experience with them and have little knowledge about what ontologies are available for reuse, even when using ontology libraries to locate them (d'Aquin and Noy, 2012). Also, if a user determines that an ontology is inadequate to the description of the information they need to work with, i.e., they need to add a property to a class, then the user needs to resolve the change in vocabulary. For example, they will more than likely need to create a new ontology using an ontology editor with which they may have little experience and make the new ontology accessible

to the Semantic Web server. For Web servers that support ontology editing, this might not prove to be an issue but not all Semantic Web server implementations provide such functionality. To manage these changes, users of the Semantic Web must be privileged to update ontologies and experienced in ontology design.

Currently, most implementations of Web servers are not, by default, offered with the ability to manage semantic content, dereference URIs to self-descriptions on the Semantic Web or for content negotiation to the ontologies that describe Semantic Web content (Drupal, 2013; Wordpress, 2013; Joomla, 2013). Since many of these systems support customizations, system administrators are able to implement features that align more with how Semantic Web servers expose information. Some Web environments such as semantic wikis inherently share semi-structured content on Web pages (Leuf and Cunningham, 2001) while others require administrative setup changes in order to provide a more detailed semantic description (Corlosquet et al., 2009). Approaches to exposing RDF versions of URL documents are found in the RDF/XML specification (Beckett, 2004). Semantically enabled Web client tools, e.g., Semantic Web browsers or software agents, can request RDF if available from Web pages and if not available can alternately extract RDF from URLs. Efforts to export Web content for Semantic Web include exposing semi-structured or structured markup such as XML, RDFa and other embedded semantics (microformats) (Soylu and Causmaecker, 2009) within Web pages, exposing meaning of a URI through Web page markup. If a URL has embedded semantics, then software agents can extract information from a URL and map it directly to RDF triples. Rdfizers are software agents or APIs that can read a URL or other document and aim to systematically map non-RDF data to a meaningful representation in RDF. Several RDFizers are available (SemanticWeb.org, 2013) to help Semantic Web-based clients extract RDF from Web pages. These tools are often specialized to certain file types such as 'pdf' or 'xls', semistructured formats such as 'html' or 'xml' or specific URL pages.

Most Semantic Web servers store the semantic information managed on the system in triplestores, repositories for RDF or OWL triples (Openlink Software, 2013; Arc2, 2013; On-

totext AD, 2013). Semantic Web servers can share semantic content through self-describing URIs, through Semantic Web services (Martin and Domingue, 2007) and through SPARQL endpoints (Quilitz and Leser, 2008). SPARQL is a query language and protocol for RDF triples sent over the HTTP protocol (Prud'hommeaux and Seaborne, 2008). SPARQL enables access to all triples in a triplestore, such that machines can explore the triples to determine the vocabulary, concepts, properties and individuals described in a triplestore. Initially, SPARQL was only able to support queries and retrieval of triples but has since been enhanced to update triples as well (Gearon et al., 2013). Typically, a Web server maintains a system-wide repository for the semantic information it manages. As a result, system-wide triplestores consist of triples from various ontologies and individuals created by the users and tools of a server. Client tools access SPARQL endpoints to search for and download structured data, as triples. In addition, some tools collect triples from different sources to process it and then might share the processed information in another SPARQL endpoint. For example, Sindice (Oren et al., 2008) is a semantic-based URI index tool that builds triples for indexing URIs and text found in RDF documents to load into a Virtuoso triplestore (Openlink Software, 2013). Users can then leverage the Sindice interface to search the loaded triples and find URIs, integrate semantic information from different URIs and analyze the content loaded in the Sindice triplestore. Since Sindice shares the information in its triplestore through a SPARQL endpoint, another client tool can subsequently query Sindice for a specific URI or set of URIs. Client tools can query multiple SPARQL endpoints as federated queries (Quilitz and Leser, 2008) and utilize information integration techniques to align data from separate SPARQL endpoints (Noy, 2004).

2.4 Linked Data

Linked Data is an approach to address the distributed nature of data holdings on the Web with the ability to link data through self-describing URIs in an effort to make the Web more understandable to machines through meaningful links (Berners-Lee, 2010). Linked data

principles include: naming things with HTTP URIs so they are Web-accessible, providing meaningful descriptions for URIs using RDF or other structured languages and linking to other URIs so related information can be discovered (Bizer et al., 2009a). Linked Data was initially focused on linking open data, where data is placed and self-described openly on the Web and Semantic Web. In such an environment, machines can meaningfully traverse links to related data without concerns for passwords, varied protocols, or other access hindering details. Organizations have worked extensively to create and link semantic data. DbPedia has reproduced a large portion of Wikipedia as linked open data (DbPedia.org, 2013), bio2RDF is an effort to publish interlinked life science data to support biological knowledge discovery and data reuse (Belleau et al., 2008) and PublishMyData is a Semantic Web portal enabling users to upload and link structured data from databases or Excel spreadsheets (PublishMyData, 2013). Linked data has helped fulfill an important piece of the Semantic Web vision by exposing structured collections of information and linking them into a single global Web. Thus far, the linked data cloud is functioning as a useful infrastructure for a growing, yet comparatively small (to the Web in general), set of users (Bizer et al., 2011). There are growing efforts to educate users in creating linked data (Heath and Bizer, 2011; Bizer et al., 2008), assess the quality of linked data (LOD-Around-The-Clock, 2012; Berners-Lee, 2009) and encourage more participation with linked data by reducing some of the initial Semantic Web challenges such as open data and ontological commitments to allow for new data licensing and open dataspaces that leverage a pay-as-you-go methodology for shared and integrating structured data (Heath, 2009; Bizer, 2010).

2.5 An Illustration of the Linked Data Principles

Creating linked data is based on four principles. These principles describe a standard approach for identifying resources on the Web and meaningful connections between them, i.e., linking content from different sources. Basically, linked data approaches can be used

to link anything that is described on the Semantic Web; linking data can be focused on the details of linking openly shared scientific data or for linking abstract concepts such as people, events and publications that are described on the Semantic Web. One benefit to using the Semantic Web and Linked Data approaches is the result global dataspace that many tools, and thus users, can rely on to understand the Web. The dataspace is built from the different resources described on the Semantic Web and the links or meaningful relationships between them. The Linked Data dataspace can be created parallel to any existing efforts to integrate or exchange information published on the Web. The section discusses Linked Data principles in more detail.

Figure 2.5 illustrates the Linked Data dataspace with many resources that might be produced by scientists, as well as images for scientists themselves; there are people, posters, programs and datasets in the diagram. The figure is an example of the Web as it primarily exists today, where Web resources might have a semantic description and most have little to no explicit links between each other, in particular, if they are stored at different data sources.

Linked Data Principle 1 requests that Web resources be identified through HTTP URIs so that humans and machines have access to them. Some resources in the diagram have green boxes, reflecting a URI using the HTTP protocol. A few do not, reflecting those that might not ever have a Web presence.

Each URI identified in the green boxes are dereferenceable complying with Linked Data Principle 2. Resources can be stored at a Semantic Web server that exposes semantic descriptions about the resources via HTTP and might also expose the resource itself or may employ a secure protocol so the resource can be downloaded. URIs that are referenced but have no Web presence, e.g., something that has no description on the Web, would be an example not complying with Principle 2. This is the case if, for example, a reference is made to a person through a fake URI that is not actually defined at a server, i.e., when the URI is requested through a browser there is no information about it. For a single URL, there may be multiple URIs in a semantic description. When Linked Data Principle 2 refers

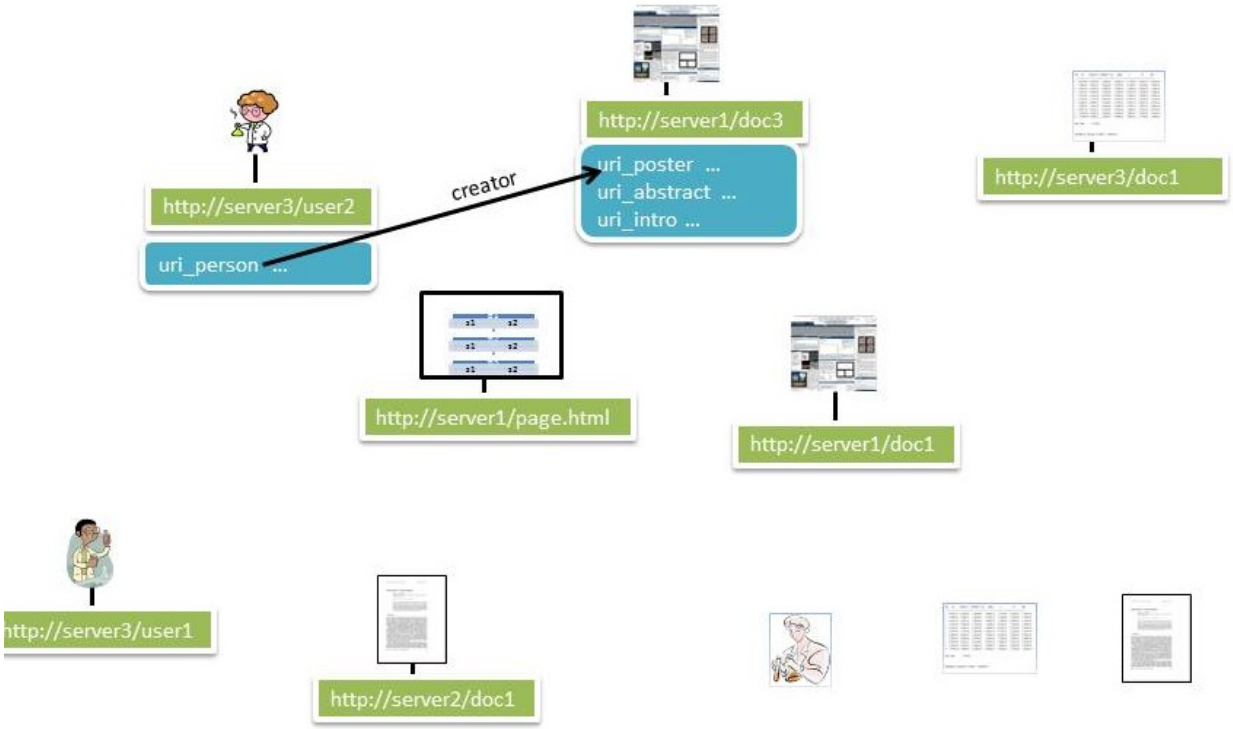

Figure 2.5: A diagram representing scientific resources. Some comply with Linked Data principles others are not even Web accessible.

to URIs being HTTP accessible, the principle refers to URIs at this level of detail.

Some of the resources in the diagram have a blue box, depicting semantic descriptions for Web-accessible URIs. Notice that some of the Web accessible resources do not have semantic descriptions. This is the case today on the Web where URIs are dereferenceable to a Web page but do not have semantic descriptions. Linked Data Principle 3 assures that when a Web server receives a request for a self-describing URI, the server dereferences the URI to a meaningful description in a structured language, e.g., RDF. RDF provides the uniform representation to consistently process information across the global Linked Data dataspace. Some URIs are dereferenceable to semantic descriptions only, i.e., the content negotiation for the resource only returns a structured representation, not a human readable representation such as a Web page.

Finally, some of the URIs in Figure 2.5 are linked, with named links to other URIs.

Linked Data Principle 4 requests that URIs have relationships that specifically link resources at the data level, not just through hyperlinks, i.e., unnamed links at the document level. The different namespaces, i.e., http://server1 vs. http://server2, imply that links can be specified across data sources.

2.6 Global Linked Data Dataspace

The Web has changed how information is accessed and managed. Where before applications expected all managed information to be stored in a single database, e.g., a relational database, the Web requires that applications learn to manage and access large amounts of data from various sources. Dataspaces are loosely connected data sources that function separately to acquire, manage and share resources, but work collaboratively to provide base functionality, for example to search, and increase integration efforts as needed (pay-as-you-go) (Das Sarma et al., 2008). In particular, dataspaces aim to address information management issues across data sources such as search and query, integration constraints, consistent naming conventions and evolution of data and metadata.

The Semantic Web and Linked Data help mitigate difficulties in working with distributed and heterogeneous information on the Web, first by assuring that information that is heterogeneous on the Web is described uniformly and second by linking related information to create a web of information. Thus, the Semantic Web and Linked Data practices provide features for processing the distributed and heterogeneous holdings of the Web so that software agents can work across a global dataspace. The Linked Open Data Cloud (LOD Cloud), shown in Figure 2.6[3] shows an example. Each circle on the diagram represents a provider of semantically described data and the links between circles represent the published links to other data providers, both components of the LOD Cloud. Providing access to the global dataspace where information can be processed and meaningfully related to other Web resources allows applications to operate on top of an unbounded set of data

[3]Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lodcloud.net/

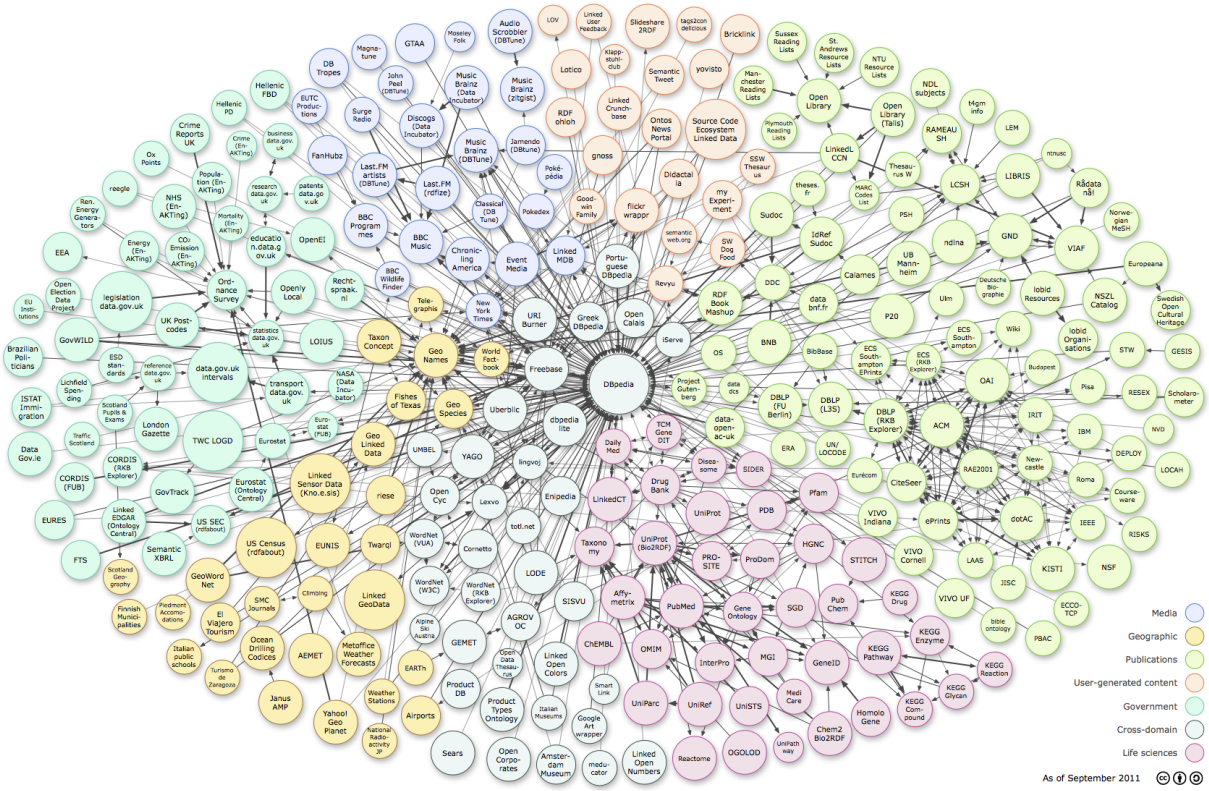

Figure 2.6: A representation of the Lined Open Data Cloud from 2011. The nodes represent organizations that have shared data on the Linked Data Cloud using Linked Data principles.

sources using uniform access mechanisms (Bizer et al., 2009b). Nevertheless, to make use of the global dataspace enabled by the LOD Cloud and Linked Data in general, humans and software agents must still locate and contextualize related information that might be distributed across the Semantic Web.

By assuring that only URIs that are dereferenceable on the Semantic Web are used for Linked Data, the Linked Data principles create a global dataspace that is accessible using a similar protocol and exposing data using a similar structure. As a result, consumers can consistently access structured representations of resources and follow links to understand more about the Web. Consumers do not have to use different or proprietary APIs to understand the resources across the Web and they do not have to be able to open and

process the resources to understand what they are. Moreover, if a dataset is completely exposed as Linked Data, i.e., each cell in each row of a data table is described as RDF triples, machines can process the Linked Data on the global dataspace along with all other information on the dataspace. Such are the holdings of the LOD Cloud, shown in Figure 2.6. It is up to information managers, however, to decide the level of detail to expose data, offering summaries when a URI is dereferenced to describe the meaning and relationships of resources in lieu of dichotomizing datasets into RDF triples, should such a practice be of concern (Gándara and Lapp, 2012).

For scientists that share their research distributed on the Web, enabling others to find and understand the information might only be humanly possible through searching data providers. Current efforts in sharing and using information over the Semantic Web have resulted in a variety of tools, techniques, suggestions and assessments (W3C, 2012; Heath and Bizer, 2011; LOD-Around-The-Clock, 2012) to create a global Linked Data dataspace of scientific resources.

Chapter 3

The Approach

The Linked Data principles, introduced in the previous chapter, list the qualities of producing and consuming semantically structured and linked information over the Semantic Web to create a Linked Data dataspace. This chapter discusses the principles that govern the systematic process to create scientific collections of semantically structured and linked Web resources as documentation of scientific results.

3.1 Guiding Principles

In an effort to share distributed and heterogeneous scientific results in a more meaningful context, than distributed and heterogeneous over the Web, CARP describes and shares the results of a research effort as a single scientific collection. Scientific collections are meaningfully structured such that Web resources are directly accessible and queryable. Moreover, given the possibility that Web resources already have user supplied metadata on the Web and knowing that some scientists do not work with Semantic Web infrastructures to share research results, the methodology aims to reuse what metadata is available, enables scientists to express more meaning through specifying relationships and works with minimal or default information if no additional metadata is available. To support these characteristics, the following five guiding principles of CARP are:

Principle 1: Capture and describe Web resources as they relate to the collection

Principle 2: Reuse existing information and structure from Web resources

Principle 3: Employ defaults to facilitate automation when capturing and structuring related information

Principle 4: Explicitly capture relationships and rules to describe the resources of a collection

Principle 5: Provide mechanisms for machines to access resources and resource information in a scientific collection both entirely and selectively over the Web

The principles are compatible with the Semantic Web vision and align with Linked Data principles. More importantly, the principles guide the methodology in describing the results of scientific research efforts.

3.1.1 Principle 1: Relate resources to a collection

Principle 1: Capture and describe Web resources as they relate to the collection

Principle 1 emphasizes the need to stay focused on the topic of the collection. The topic of the collection is the reason resources are related to each other. For example, a scientist may have the URIs for a publication on joint inversion, a dataset resulting from a joint inversion study and a program that was used to conduct the joint inversion study. Principle 1 recommends staying focused by adding resource descriptions in a *bottom up* approach to the collection. The term *bottom up* is used to mean that resources are added to a collection regardless of how they relate to other things in the collection or on the Web. In addition, because scientific collections collect Web resources, each resource is Web accessible through its identifier (URI) in the collection.

Figure 3.1 shows the resources from Figure 2.5 with some resources added to a scientific collection. CARP Principle 1 expects all Web resources that are part of a collection to be explicitly identified. Figure 3.1 shows a gray box that includes the resources that have been added to the collection. Initially, at least, the URL for documents that are on the Web are added.

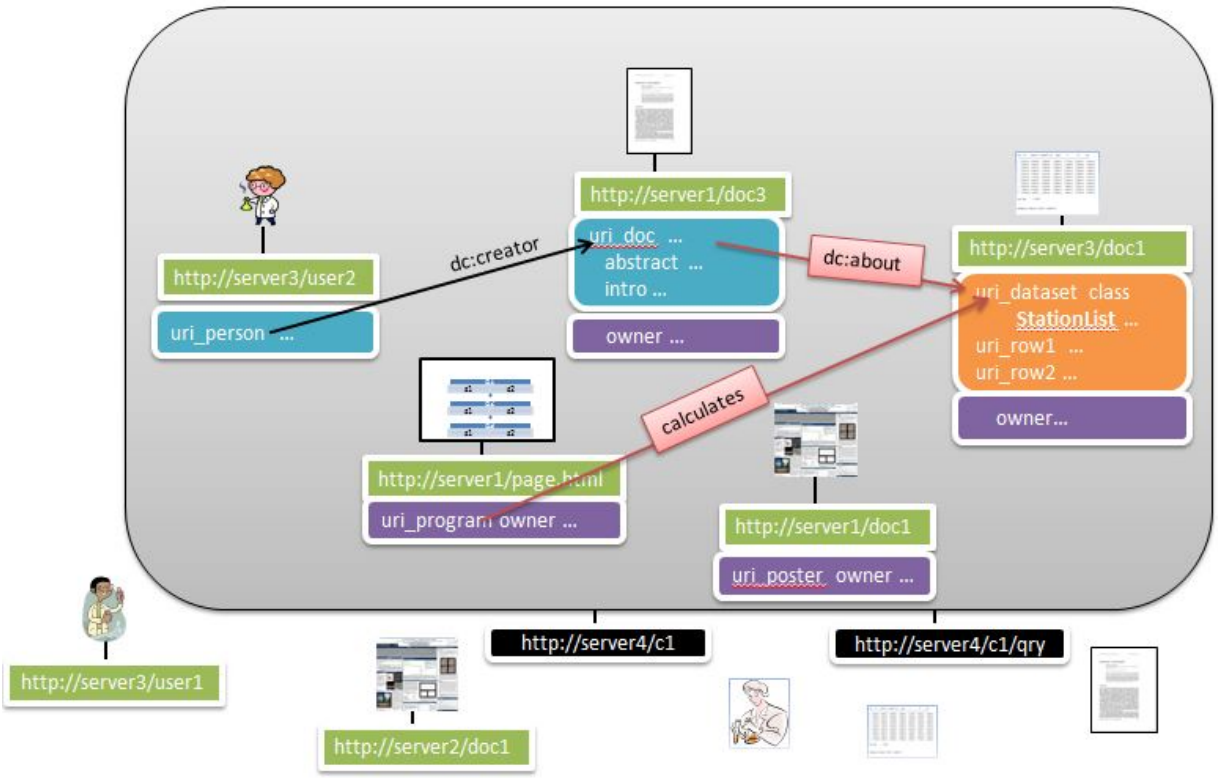

Figure 3.1: The resources in a scientific collection reuse existing Web information, add additional information to describe resources, and add relationships between resources.

3.1.2 Principle 2: Reuse existing Web content

Principle 2: Reuse existing information and structure from Web resources

Principle 2 emphasizes that information should not be redefined, in particular, if there is already an authentic semantic description available. The most accurate and authentic description of a resource should come from the server that dereferences the resource's URI thus a collection should obtain self-describing information about a URI when available. For example, resources that are uploaded to a data portal often have accompanying metadata to describe the resource. If this information is not semantically structured, it should be manually or automatically extracted to describe the resource within the collection. Note that

an authentic description of a resource may not be initially available from the server dereferencing a URI, however, if made available, the authentic description should be obtained. This is another example of how semantically described collections are created *bottom up*; there is no semantic vocabulary imposed on the resources added to a collection.

The publication resource shown in Figure 3.1 has semantic properties defined at the server that dereferences it, these are imported as part of the description in the collection. The same is done for other Web resources. As a result, semantically described collections are uniquely described from the terms of the resources added and the collections evolve as more resources are added and authentic descriptions are imported.

CARP Principle 2 expects to reuse information already shared on the Web about Web resources, in particular, if the resources are already dereferenceable to a structured description. However, not everything will have semantic descriptions thus extracting RDF or RDFizers are acceptable approaches to obtaining more information about a Web resource. Figure 3.1 shows a dataset with an orange box representing a table that was RDFized into a semantic description about the dataset.

3.1.3 Principle 3: Employ defaults

> *Principle 3: Employ defaults to facilitate automation when capturing and structuring related information*

Principle 3 is meant to minimize the decisions and distractions that are required to create semantically described collections, by capturing minimal information initially and adding specifics as they are known. At a basic level, scientific collections should identify resources, i.e., URIsa and, optionally, any default properties added to resources in the collection. For example, the Semantic Web community has already established vocabularies, from ontologies, that are designed to express common attributes like location, people, creator, and more (Brickley and Miller, 2010; Brickley, 2006; Weibel et al., 1998). Principle 3 advises

that attributes about resources should be captured and mapped to common semantic vocabularies, especially when no mapping to a semantic vocabulary can be obtained for a Web resource.

Knowing that the Semantic Web is still growing and that some resources may never be supported with semantic descriptions when they are dereferenced and content is returned, CARP Principle 3 requests that defaults be provided and that common Semantic Web vocabulary or local namespaces be used to facilitate adding structured information. One way to employ defaults is by adding default attributes to each resource concerning licensing, ownership, creators, etc. and mapping the attributes to related Dublin Core properties. Figure 3.1 shows that all resources in the collection have properties in purple boxes which are default properties consistently added to resources, whether they have an existing semantic description or not.

Another way to employ defaults is by enabling for properties or classes to be created in a default collection namespace, to avoid the task of creating new ontologies or extending a published ontology to support new terms. In addition, creating new vocabulary in a default collection namespace avoids the distraction of searching the Semantic Web for vocabulary that has little relevance to a scientific collection or requiring that a separate tool, e.g., an ontology editor, be used to modify an ontology. The dataset object identified in the orange box of Figure 3.1 is identified as an individual of the **StationList** class, a class declared in the local collection namespace.

3.1.4 Principle 4: Capture explicit relationships

> *Principle 4: Explicitly capture relationships and rules to describe the resources of a collection*

Principle 4 is meant to emphasize the need to support explicitly capturing relationships between resources. CARP Principle 4 requests that relationships between resources, either in or out of the collection, be explicitly specified in the collection. Principle 4 relies on the

knowledge built from the previous three principles by enabling links to be added between URIs identified from RDFizers, self-described URIs or extracting RDF and by adding new relationship names to the local namespace if no other vocabulary is known.

Figure 3.1 depicts relationships between URIs, added to a scientific collection, as red arrows with the name of the relationship in a red box. Notice that the **about** relationship is from the **Dublin Core** vocabulary. **Dublin Core** is an established vocabulary with terms that describe common things and is one example of a common vocabulary that facilitate describing Web resources. The **calculate** relationship is added to the local namespace.

Social annotations, such as comments and discussions are another example of when explicit relationships can be captured (Gándara et al., 2011a; Bowers and Ludäscher, 2006). For example, if a comment is captured about a resource in a collection, the comment should be explicitly related to the resource's URI, not just added to the collection.

In addition, rules are often used to describe semantic information. These should be captured and added to the collection, if they are relevant to the scientific collection. Further consideration for rules and reasoning within the CARP principles is future work for the research.

3.1.5 Principle 5: Enable machines Web access

> *Principle 5: Provide mechanisms for machines to access resources and resource information in a scientific collection both entirely and selectively over the Web*

Principle 5 emphasizes providing access to all of the information documented in the semantically described collection. If the information is shared entirely, including the vocabulary and data, then client tools can adapt to the uniqueness of a collection within the context of the client tool. Figure 3.1 shows two black boxes, one for a URI that exposes a serialized collection, i.e., a self-describing URI about the collection, and a query URI to enable searches and access to specific resources along with their meaning and relationships, based on a query.

Notice that principle 5 does not advocate for open access to information that should be private or secure. Principle 5 still functions in a secure environment and under the constraints of user privacy. Privacy is discussed in future work for the research.

Principles 1 to 5 assure qualities in creating scientific collections such that research efforts can be described leveraging information from the Web and Semantic Web. As a result of following CARP principles, scientific collections can be created to enhance the global Linked Data dataspace.

3.2 Significance of Guiding Principles

This section discusses the progression of the Web to a global Linked Data dataspace, then focus on how using the Linked Data dataspace reduces the focus of processing Web content to such issues as consistent URI naming, information integration and exchange, and enhancing Linked Data. In addition, this section discusses the role CARP principles play in enhancing the Linked Data dataspace.

3.2.1 A Structured Information Base

The Web itself was established to create a large information base (Jacobs and Walsh, 2004). Initial principles, good practices and constraints concerning Web content are focused on general availability of Web documents with unique addressing. Web principles provide suggestions for URIs, such as the use of URIs to unambiguously access descriptions and representations of Web resources, assuring the quality of URIs and assuring the persistence of URIs for continued access over the Web. Additional principles are suggested to protect users from side effects of referencing or dereferencing resources and expecting orthogonality of Web concepts, e.g., a URI should not require modification if the representation of a resource referenced changes, that allow the Web to continuously change with limited referential effects on agents using Web content. These basic Web principles promote growth of a distributed and heterogeneous Web of information. Consequently, the ability to integrate

and exchange distributed Web content requires inter-organizational agreements of APIs, services and data structures.

Where the principles for the Web mainly concentrate on access and traversing of Web documents, the Semantic Web focuses on meaningful descriptions and integration of data from diverse sources. Principles supporting the Semantic Web function over the principles of the Web, for example, by employing the use of URIs to access Web resources. In addition, Semantic Web principles include support for naming types and links of information on the Web (Koivunen and Miller, 2001) to facilitate information exchange and integration. Semantic Web principles suggest more support for a growing Web by tolerating partial information that evolves as more structure (types) is shared. Additionally Semantic Web principles emphasize trust based on individual software agent evaluation, so processing the Web can avoid unnecessary validation or a centralized verification. Another principle of the Semantic Web is to minimally impose standards and only as they are necessary, enabling the content and structure of the Semantic Web to grow based on use. Through Semantic Web principles, information shared on the Web can be shared based on structured types, enabling machines to access a more meaningful understanding about Web content. As a result, integration and exchange is based on uniform structures that can be communicated through APIs, services and other Web-based protocols.

Linked Data principles are meant to extend Semantic Web principles assuring that Web content can be consistently processed. Linked Data principles focus on creating typed links to things, not just documents. These principles rely on Web servers to expose structured and uniquely identified information, via HTTP, that is related to other structured and uniquely identified information. The use of URIs to consistently identify Web-accessible resources assures that the meaning of the those resources can be obtained consistently. As a result, software clients use the information on the Linked Data dataspace as a distributed yet consistently accessible (through HTTP) knowledge base. Tools can reuse this information and enhance it by exposing additional Linked Data that adds properties and relationships about URIs on the Linked Data dataspace. Enhancing the Linked Data

dataspace can only be done, however, if tools share information using the *same URIs*, not through tool-specific representations or identifiers, and by sharing information *consistently structured and linked*, not in a proprietary format or through tool specific APIs or services. By sharing information in RDF, for example, all triples are consistently structured, despite specific vocabulary that might be used to characterize Web objects. Notice that the Linked Data dataspace can exist in conjunction with other techniques for sharing information through Web servers, information managers and users do not have to abandon existing data sharing practices to add information to the Semantic Web or Linked Data dataspace. Linked Data principles provide a foundation for sharing information consistently so that other tools can rely on that information for processing Web content.

3.2.2 Issues

If the Linked Data principles are followed to produce and consume resources on the Web, then tools accessing those resources expect to find meaningful information, i.e., structured and linked data, through dereferencing URIs using the Web's http protocol. As a result, data providers and consumers can concern themselves with a reduced set of issues, e.g., URI naming standards, agreeing on vocabulary to structure and exchange information, and enhancing linked data on the Linked Data dataspace. The remainder of this section elaborates a bit on the three examples.

Linked Data URIs. Since the Web is decentralized, it would be difficult to dictate the entire syntax of a URI. URI providers can use consistent guidelines, best practices or random but unique names (Sauermann and Cyganiak, 2008), however, adherence to URI naming rules is mostly based on individual information management approaches. When information to be referenced through a URI are database records from a database, there is a need to assure that records are consistently accessible and described. The Banff Manifesto[1] aims to provide cross-database referencing with common identifiers to the Semantic Web. The Banff Manifesto are six 'rules of thumb' that function over the Linked Data

[1] $http://sourceforge.net/apps/mediawiki/bio2rdf/index.php?title=Banff\_Manifesto$

dataspace. The goal is to focus on uniformity of reference over a 'bioinformatics semantic web' assuring that URIs are consistently and reliably named and described. For example, URIs are normalized, dereferenceable, use public namespaces and mandatory predicates. In addition, there are expectations in quality of meaningful descriptions for resources such as publicly available RDFizers, dereferenceable ontologies and avoiding blank nodes. The Banff Manifesto aims to assure that accessing a URI produced by the described scheme will be consistent.

Structuring and Exchanging Linked Data. Initial efforts for working on the Semantic Web require agreements to assure interoperability across data sources. When data providers share information on the Semantic Web they are usually required to use vocabulary, either from another source or created by themselves, if there are hopes to integrate or exchange this information with others. This is the case with Linked Data as well, however, using pay-as-you-go principles enables less of an upfront effort to publishing data on the Linked Data dataspace (Bizer, 2010). Pay-as-you-go for Linked Data suggests that common vocabulary be used, e.g., Dublin Core, Foaf, and that concerns for integration with other objects on the Semantic Web be ignored until they are necessary. At the point where an integration is needed, both the producer and consumer can negotiate a vocabulary alignment. The vocabulary alignment can be published on the Web for other software agents to reuse. Using this approach enables the Linked Data dataspace to grow without concerns about integration, knowing that data integration can be handled as a distributed effort between data producer, consumer and other parties when needed. Moreover, if an existing integration exists, it can be reused.

Enhancing the Linked Data Dataspace. The significance of the CARP principles is in addressing the issue of enhancing the Linked Data dataspace. Tools that implement CARP assure that scientific collections reuse, preserve and expose Web information so that the collection is compatible with the Linked Data dataspace.

The CARP principles describe qualities for consuming and producing enhanced linked data. The alignment of CARP with the Linked Data dataspace will be illustrated with

Figure 3.2, a variation of Figure 3.1 that highlights this discussion. The orange boxes list the principles.

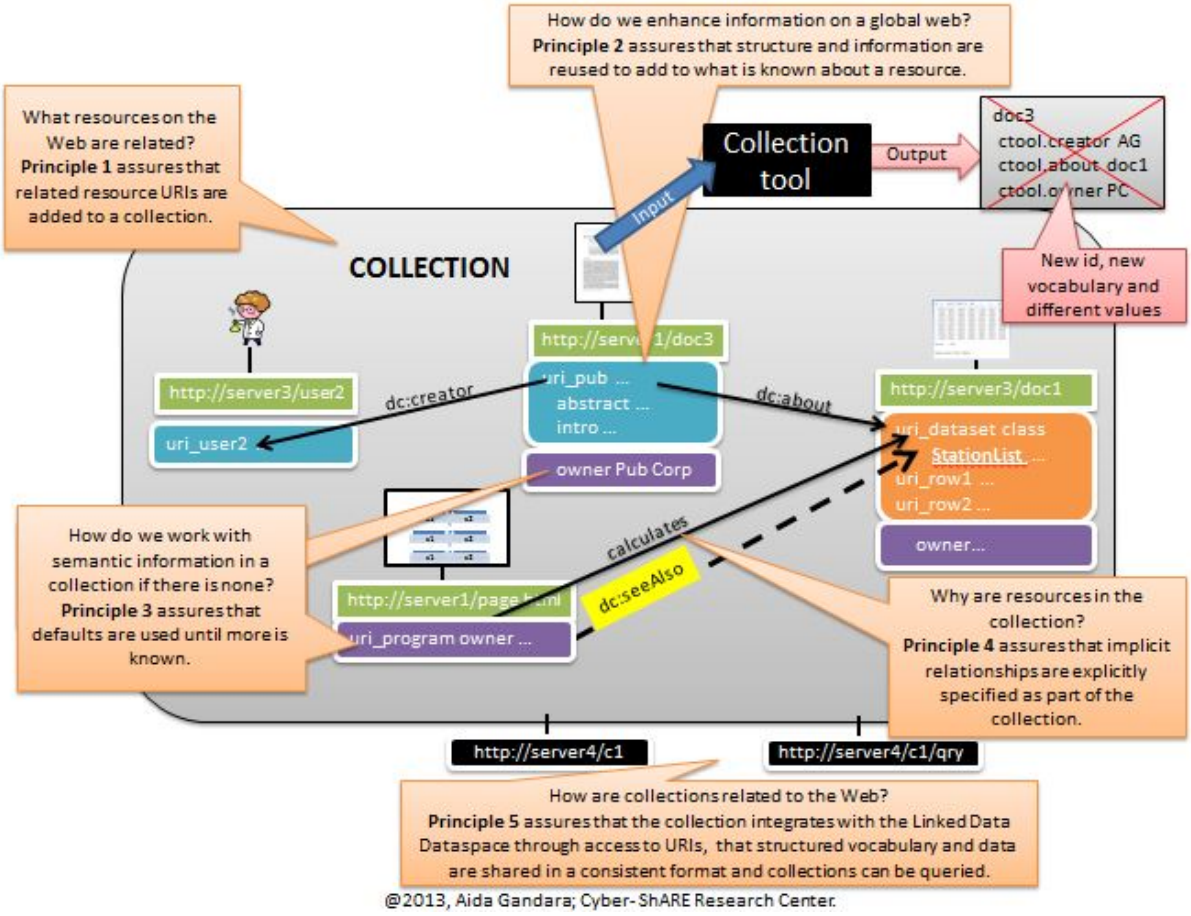

Figure 3.2: CARP principles assure that the creation of collections of information on the Web reuse, preserve and expose structured information so it is compatible with the Linked Data dataspace.

The first CARP principle, *capture and describe Web resources as they relate to the collection*, initiates the process of creating a scientific collection. As tools build a scientific collection, one goal is to identify those Web resources that are directly related to the collection, versus those that are available on the Web that have no relation. The linked open data cloud, described in Section 2.6, shows a growing Web of information. Following links is one approach to finding related information and, as seen over many years of following

hyperlinks, a useful one. Sharing related information as an interlinked collection, however, facilitates identifying the simple fact that Web resources are related in a specific context. Principle 1 *provides a starting point for the resources that will be described in a collection*, as opposed to resources found over the entire Web or the Linked Data dataspace. In addition, Principle 1 *preserves the Web URI of resources assuring that resources in the collection remain related to the Web.* Without following Principle 1, identifying the resources that are part of a scientific collection is unclear and must be derived from following links, if they exist, on the Web or Semantic Web.

The second CARP principle, *reuse existing information and structure from Web resources*, reuses the Linked Data dataspace. Tools that work with information from the Web by importing that information facilitate knowledge gathering. Following this principle *facilitates reuse of meaningful information in describing resources*, reducing rework by those building the collection. Without following Principle 2, information that has been provided on the Web is either ignored or re-entered. Moreover, in order to enhance knowledge about existing Web resources, the structure and vocabulary should be preserved so that knowledge about the resource is consistent. Figure 3.2 shows an example. The diagram shows a tool, as a black box called Collection tool, that works with information on the Web. If a tool references a resource and does not reuse meaningful information about the resource, then a new description about a resource may be introduced on the Web with little to no relation to what was already there, e.g., new object identifier, new vocabulary, different values for attributes. This is the Output of the Collection tool.

The third principle, *employ defaults to facilitate automation when capturing and structuring related information*, establishes a basic environment to work with, in case there is little known or available about a resource initially. Tools that work with information on the Web have expectations in order to process information. Working in a default environment, e.g., requiring little initial knowledge, and allowing for information to be added as it is known facilitates enhancing the information, because users can start working with tools and data without a lot of setup. The Semantic Web's growth has occurred over the last 12

years and is expected to continue, thus, additional semantic descriptions are expected to increase. Tools must adapt to this fact in order for the Semantic Web to grow and enable processing of heterogeneous and distributed information as a global dataspace. Without this principle, resources such as the poster or the program in Figure 3.2 would not have additional properties to semantically describe them and may not be added to a relevant collection. In addition, not supporting locally defined vocabulary means that scientists must understand how to edit vocabulary using other tools. This would be a distraction while documenting scientific research.

The fourth principle, *explicitly capture relationships and rules to describe the resources of a collection*, assures that collections provide links between resources, i.e., URI to URI, as part of the definition of the collection. Creating a collection of related resources is a start to understanding how things are related, e.g., that a set of resources are all related to a scientific research effort, but enabling scientists to add implicit relationships explicitly adds more meaning. Figure 3.2 shows a link between two resources using the **Dublin Core seeAlso** relationship (in a yellow box and dotted arrow). This is a generic reference relating two resources and would be helpful to knowing that they are related. With the **calculates** relationship, additional knowledge is conveyed to understand how the program and dataset are related. Without Principle 4, the resources found in a collection are related simply by their existence in the collection. Adding explicit relationships provides more meaning about the resources and why they are relevant to the collection. Supporting the capture of explicit relationships in one environment, versus having to use different tools or environments to add resources, vocabulary and relationships facilitates building a meaningful collection.

The fifth principle, *provide mechanisms for machines to access resources and resource information in a scientific collection both entirely and selectively over the Web*, adds the collection to the Linked Data dataspace by preserving URIs in the collection and exposing all information added to the collection through queries. As the Web is today, organizations must find the best approach to integrate and exchange information with other organizations and are often limited in scope, e.g., to a group of organizations and specific domains of

data. By exposing everything about a collection, i.e., the data and the structure, in a uniform format, access and integration of collections becomes more feasible for machines that need to process the heterogeneous and distributed information currently on the Web.

3.3 The CARP Methodology

Current efforts in sharing and using information over the Semantic Web have resulted in a variety of tools, techniques, suggestions and assessments to create, manipulate and consume semantically described information. There is no systematic approach to documenting scientific research or sharing it on the Semantic Web. This chapter describes the methodology identified for this research. The methodology is called CARP, which stands for the **C**ollect-**A**nnotate-**R**efine-**P**ublish Methodology. The methodology shares documentation about scientific research over the Semantic Web. CARP describes scientific research results as a scientific collection of Web resources such that the Web resources are semantically described using existing Web content, relationships between Web resources are explicitly identified and the collection is searchable over the Semantic Web.

3.3.1 Overview

CARP is the result of enhancing previous efforts to document scientific research related results (ci1, 2010; Gándara et al., 2011b; ci1, 2012). The initial approach worked with non-semantic representations and was altered to support Semantic Webs and Linked Data techniques. Initial work and its relevance to this final methodology are discussed in Appendix A. One finding from previous efforts was a need for systematic steps to document research efforts by describing scientific results. Since scientists are publishing work on the Web it is necessary to understand how to take advantage of their efforts, in particular, when adding metadata to upload resources and to describe the relationships between the distributed resources shared. Figure 3.3 is a diagram depicting the four phases of CARP.

The diagram in Figure 3.3 shows each phase of the methodology in a separate oval,

around a collection of related information, i.e., a scientific collection. Three phases, **C**ollect, **A**nnotate and **R**efine, are responsible for adding more meaning to the collection, hence these each have an arrow pointing into the collection. The role of **P**ublish is the dissemination of information from the collection, hence, the arrow pointing out. **C**ollect is the start phase for any resource to be included in a collection. A resource will only be included in the collection once, for this reason there is only an arrow pointing into the collection from this phase. **A**nnotate and **R**efine are dependent on the information in the collection and can be continuously performed to enhance the meaning of resources within the collection; this is depicted with the arrows leading from the collection to these phases and back. **A**nnotate and **R**efine require access to URIs so that relationships can be explicit. **P**ublish shares the information in the collection on the Web and Semantic Web; the arrow from the collection to this phase represents the role **P**ublish has in exposing the contents of a scientific collection. The remainder of this section discusses each phase and how each affects the representation of a scientific collection.

3.3.2 CARP Phases

A scientific collection is a knowledge base describing the resources used and created from a research effort and the relationships between them. The knowledge about a scientific collection is stored in a description logic system, a representation for formally describing information and commonly used to describe information on the Semantic Web (Baader et al., 2010)[2]. This chapter uses a basic description logic to describe scientific collections. Mainly this research focuses on knowledge representation and information integration to describe the content of a scientific collection, however, description logics also represent reasoning. Reasoning will be considered in future work for this research.

A scientific collection is denoted as a pair $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, where $\mathcal{T}$ introduces the terminology, i.e., the vocabulary of the knowledge base and $\mathcal{A}$ defines assertions about named

[2](Baader et al., 2010) describes description logics to represent the terminology (TBox) and assertions about individuals (ABox) of a knowledge base.

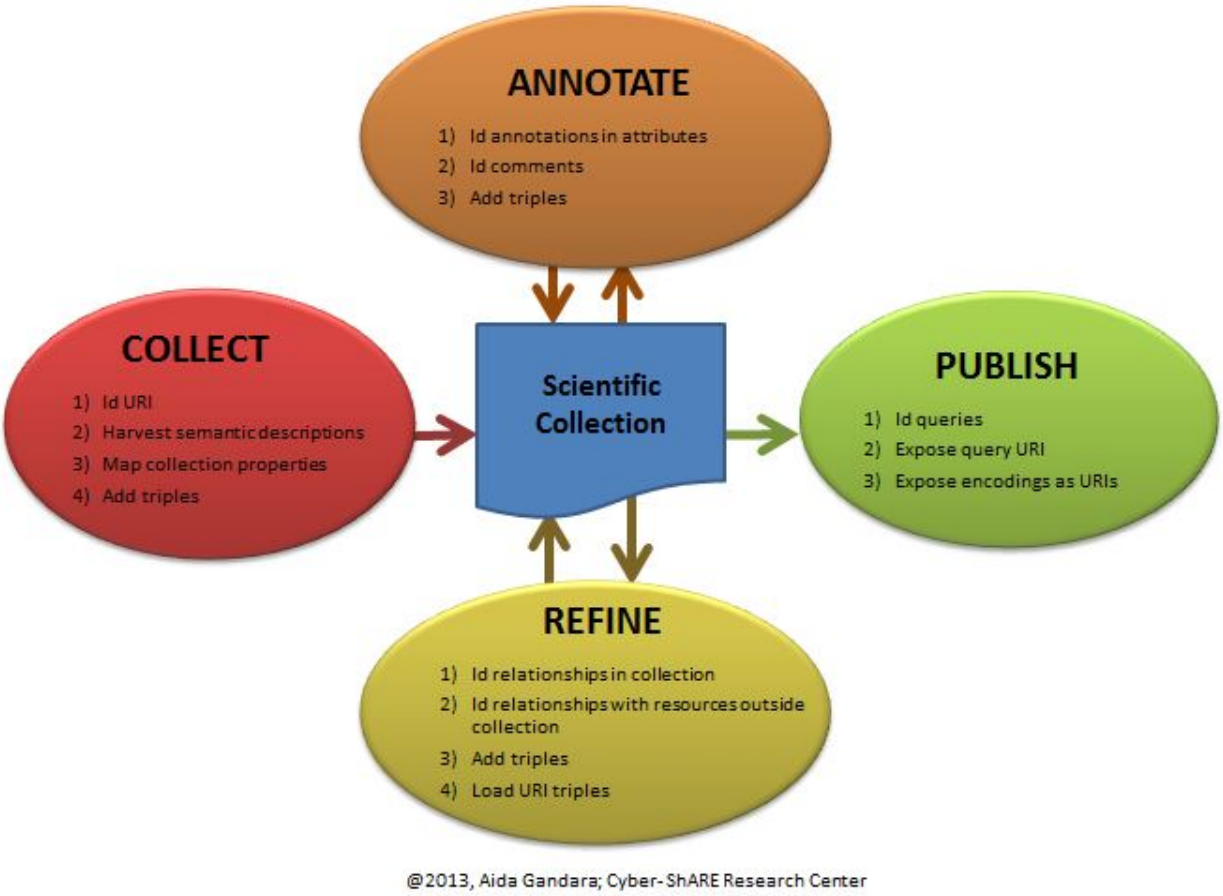

Figure 3.3: The four phases of CARP: Collect, Annotate, Refine and Publish, each shown in an oval with arrows that designate the flow of information to and from a scientific collection. The scientific collection is in the middle.

individuals in terms of the vocabulary. $\mathcal{T}$ consists of concepts that describe individuals and roles that describe binary relationships between individuals. $\mathcal{A}$ describes a specific state of an application domain, in terms of concepts and roles.

To describe a scientific collection, the following constructs are introduced:

c : a constant value, such as a string or number

uri : any Web-accessible URI

$\mathbf{uri}_c$: a URI of the collection

uri$_r$: a URI of a resource in the collection

uri$_a$: a URI that resolves to a comment. For this methodology, a comment has its own Web-accessible URI.

uri$_p$: a URI that resolves to a description of a person. For this methodology, the description of a person has its own Web-accessible URI. People tend to have multiple URLs associated to them, e.g., multiple Web pages for personal use or due to their participation with an organization. This methodology expects all people to have a single self-describing URI which can include a link to the multiple URLs created for the person.

The following concepts are introduced to describe individuals in a scientific collection:

Collection : the class of objects that represent a CARP scientific collection.

Resource : The class of Web resources. These can be various classes of Web-accessible research objects, e.g., a publication, poster, etc.

Annotate : the class of objects that represent a textual annotation related to a resource.

The following roles are introduced to characterize individuals in a scientific collection:

definedBy : a property identifying an object that defines another object.

member : a property to relate one object as the member of another object.

seeAlso : a property to relate one object to another object.

previous : a property to identify that one object occurred before another object.

about : a property that identify that one object is about another object.

description : a property relating text to describe an object.

comment : a property to identify the string of a comment object.

time : a property to identify the time related to an object.

squery : a property to identify a query string that applies to a scientific collection.

queryu : a property identifying the Web location (URL) to query a scientific collection.

contributorOf : a property to specify the object that identifies the person that contributed to an object.

p : a generic term to signify properties imported from another knowledge base or terms introduced in the scientific collection's knowledge base. Such properties could be attributes, where the value is a string, or relationships, where the value is another object.
A Collection named individual (a scientific collection) is created with the following concept and role assertions:
Collection(uri_c) - the collection object
definedBy(uri_c,uri) -
description(uri_c,c)
user added attributes: **p**(uri_c,c)
and user added relationships: **p**(uri_c,uri)

The phases of the methodology are responsible for making assertions about individuals. In addition, a collection can import other knowledge bases, thus the terminology and assertions in this section are only what CARP specifically adds. As a result, imported knowledge bases can add more complex descriptions of concepts and roles. Each phase is discussed below, specifying the constructs added to a scientific collection knowledge base at each phase. The constructs added are primarily named individuals and relationships although vocabulary can be added, i.e., concepts and roles, as needed. The ability to handle more complexity in editing the terminology of a knowledge base, e.g., by adding concept constructors or role constructors, can be handled as needed to represent details of a scientific collection.

Collect

The first phase, **C**ollect is meant to identify the resources added to a collection. **C**ollect has four main functions 1) to identify the web-accessible identifier (URI) of a resource that will be added to the collection; 2) to harvest the semantic description of a resource; 3) to map any other metadata collected about a resource to structured vocabulary; and 4) to add the structured resource description to the collection.

There may be preliminary steps before adding resource URIs to a collection, including: 1) identifying ownership of the resource, licensing of its reuse, the actual location of the resource, obtaining a description and keywords. Mechanisms should be used to facilitate and guide entry of resources, e.g., by allowing users to capture details in the columns and rows of a spreadsheet; 2) uploading resources that are on local boxes to a Web server that supports self-describing URIs or that will embed metadata about the resource in an HTML page as a microformat; 3) assuring that resources have attribution to the authors; 4) obtaining URLs describing the organizations that own resources used by the research effort, preferrably one that provides semantic descriptions or embedded microformats describing the resource; and 5) consider building archives or large filesets of related files so that they can be downloaded together, to avoid describing large quantities of files separately or downloading them separately.

The scientific collection is created bottom up by adding resources and then relying on the rest of the phases to enhance the meaning with annotations and relationships. A resource is identified by a Web-accessible URI and attributes that describe it. In some cases, a Web-accessible URI is not available for a resource. If this is because the resource is not digitally available or because the resource can not be accessed directly with a URI, a description of the resource is created using default properties for the collection and a collection URI is assigned. For resources that are not Web-accessible, a description property may be added to explain how to access the resource, e.g., the Website that provides it. In cases where a digital resource is not yet shared on the Web, the resource should only be uploaded to the Web by the owner of the resource and the resource assigned a Web-accessible URI. A resource URI added to a collection in the **C**ollect phase is added as a **member** of the collection. This research assumes that, although a resource itself may not be accessible to all users that access to content of a scientific collection is at the collection level, i.e., a user has access to all or none of the collection based on privacy rules.

The semantic description of a resource comes from information found at the resource's URI and additional properties that might be collected for resources added to a collection.

If semantic information can be retrieved from a resource's URI, e.g., the resource has a self-describing URI, then this can be captured and added to the collection. Alternatively, extraction techniques may be used to extract information from a resource to set them in properties, e.g., extracting the authors of a publication and setting respective properties.

The terms that describe a resource come from the semantic description obtained from the resource's URI or from properties added to the collection ontology's namespace. If similar terms are used across many scientific collections, an ontology can be created to define the similar terms and the ontology can be imported into the scientific collection.

A Resource individual is created with the following assertions:
type(uri_r) - since a Resource can be of various types, the type assertion would vary based on a specific resource type, e.g., poster, publication.
member(uri_r,uri_c)
description(uri_r,c)
user added attribute roles: **p**(uri_r,c)
and user added relationship roles: **p**(uri_r,uri)

The **C**ollect phase aligns with Principle 1 by capturing resource information to create a scientific collection. This approach also supports reusing information that exists on the Web as identified in Principle 2, because the values of the properties are harvested from semantic descriptions of a resource or by setting properties of a resource. Finally, this phase aligns with Principle 3 by leveraging default vocabularies and a local namespace to create new ontology terms.

Annotate

The second phase, **A**nnotate is meant to identify relationships found in text, and add them as properties of a resource or through comments related to the resource. For example, if a description property for a resource states that the resource relates to a term in dbPedia, e.g.,

through a hyperlink, then a triple is added to explicitly relate the resource to the dbPedia URL. **A**nnotate has three functions 1) to identify annotations within the attributes of resources in a collection that relate to other URIs; 2) identify comments about a resource; and 3) to add the annotations and comments that form relationships between URIs to the scientific collection.

When the text in a resource's property has a link to another URI, a relationship is created between the resource URI and the referenced URI. The relationship is added to the collection as the following assertion:

seeAlso$((uri_{r1}, uri_{r2})$ where one resource is related to the other because of a hyperlink reference in the text.

This relationship defaults to the **seeAlso** property. Other properties can be applied if preferred. Properties are selected by importing the vocabulary or adding the term to the collection's knowledge base.

Allowing additional properties to be added to comments enables an annotation to be part of a larger or more robust collaboration environment. For example, SIOC, a vocabulary currently leveraged on the Semantic Web considers comments are part of online communities. Additional properties can be added to a comment to specify relationships that are part of such a specification. By default, this methodology is not describing details of an online community. However, a comment is unique and normally has attributes like date, time, and creator, of which should be captured.

A Comment individual is described through the following assertions:
Comment(uri_a1)
contributor(uri_a,uri_p)
time(uri_a,c)
previous(uri_a1,uri_a2)

contributor(uri_a1,uri_p)
about(uri_a1,uri_r)
user added attribute roles: **p**(uri$_r$,c)
and user added relationship roles: **p**(uri$_r$,uri)

The **A**nnotate phase is meant to be an optional and iterative phase; a scientific collection can be shared on the Semantic Web without annotations of resources. Relating annotations to the URIs of a resource and including annotations as part of a scientific collection aligns with Principal 1 of this methodology, using defaults when scientists have no specific ontology requirement aligns with Principle 2 and relating annotations to specific URIs aligns with Principle 4.

Refine

The next phase, **R**efine, is focused on explicitly capturing relationships about the information in a scientific collection. **R**efine, primarily has four functions: 1) to identify relationships between two resources in the collection; 2) to identify relationships between a resource in the collection and another URI not in the collection; 3) to load relationships between URIs as triples into the collection; and 4) to (optionally) load the semantic description of a URI as triples into a collection.

Relationships between URIs, whether in the collection or not, are described through the following assertions:
p((uri_{r1}, uri_{r2}))

In some cases, a URI will not have any properties describing it in the scientific collection. For example, this might be the case when a relationship is made between a URI in the collection and a URI not in the collection. During **R**efine, a semantic description for a URI can be added to a scientific collection with no limit to what is added. This is an optional step

because in some cases it may be beneficial not to load the semantic description of a URI. For example, to avoid issues when a semantic description is excessively large, inconsistent or the details of it are not relevant to the collection. Notice that loading semantic information could result in inconsistencies or contradictions, such issues should be dealt with by either modifying the scientific collection to fix it or by removing it. Further research on handing such situations are beyond the scope of this work.

By adding relationships to describe resources in scientific collections, **R**efine supports Principle 1, by reusing information already published at URIs, this phase supports Principle 2 and by capturing explicit relationships between URIs, **R**efine supports Principle 4.

Publish

The final phase of this methodology, **P**ublish, is meant to share the structured information in a scientific collection on the Web and Semantic Web. By default, a collection is accessible through a self-describing URI that contains all the triples describing the collection. **P**ublish has three main functions: 1) identify machine understandable queries that can be applied to a scientific collection and capture them as properties of the collection; 2) enable machine understandable queries to be applied manually or automatically toward a collection; and 3) expose query results as serialized encodings that are accessible through URIs on the Web. Notice, the encodings can be any human or non-human format, in order to accommodate users and client tools. **P**ublish can also be used to support encodings that would expose scientific collections for integration with other data formats or standards.

Queries are specified as properties of a collection, described with the following assertion:
squery(uri_c, c)

This methodology leverages URLs to enable machine understandable queries to be applied to a scientific collection. Support for manual and automated queries from a URL to obtain objects from a scientific collection is similar to how some triplestore management

systems currently implement SPARQL endpoints Openlink Software (2013); Arc2 (2013). For automated queries, a collection query URL is exposed on the Web such that when a client tool accesses the URL with a query as a parameter, the query is applied to the collection and the client tool receives encoded query results. To support manual queries, a collection query URL is exposed on the Web such that when a user accesses the URL with no parameters, a user interface to accept queries and display query results is displayed. Machine understandable queries should be specified syntactically compatible with the storage software for the scientific collection. The collection query URL is captured as a property of the scientific collection.

The collection query URL is described with the following assertion:
queryu$((uri_c, uri_q))$

In addition, URLs are leveraged by this methodology to expose serialized encodings resulting from queries applied to a collection. For example, if a desired encoding for a collection is an HTML Web page with the results of the query shown in a geo-spatial map, then a URL is assigned to the map Web page. A software agent would access the URL, entering a query to the scientific collection as a parameter, the query would be applied to the scientific collection and the encoded map Web page would be returned with the results showing on the map. This methodology recommends that for encodings of user interfaces, that a default query exist in case users do not enter one at the URL. Moreover, that the interface be interactive such that the query can be modified.

The **P**ublish phase aligns with Principal 5, by sharing the collection on the Semantic Web as a queryable, machine understandable collection.

Figure 3.4 shows a graph representing some of the resources and relationships for the scientific collection shown in Figure 3.1. The graph for the collection shows resources as ovals, the relationships between them are represented as black arrows with boxes labeling the relationships. As an example, properties have been added to describe the publication,

i.e., **year**, **haspublisher** and **format**. In addition, a comment has been added to illustrate annotations. The scientific collection is accessible through a query URL that can be used to explore resources and relationships as well as to answer questions about the collection. For example, for this collection, a query can be created to ask what resource **calculates** the dataset, **uri_dataset**. The query would return one resource, **uri_program**.

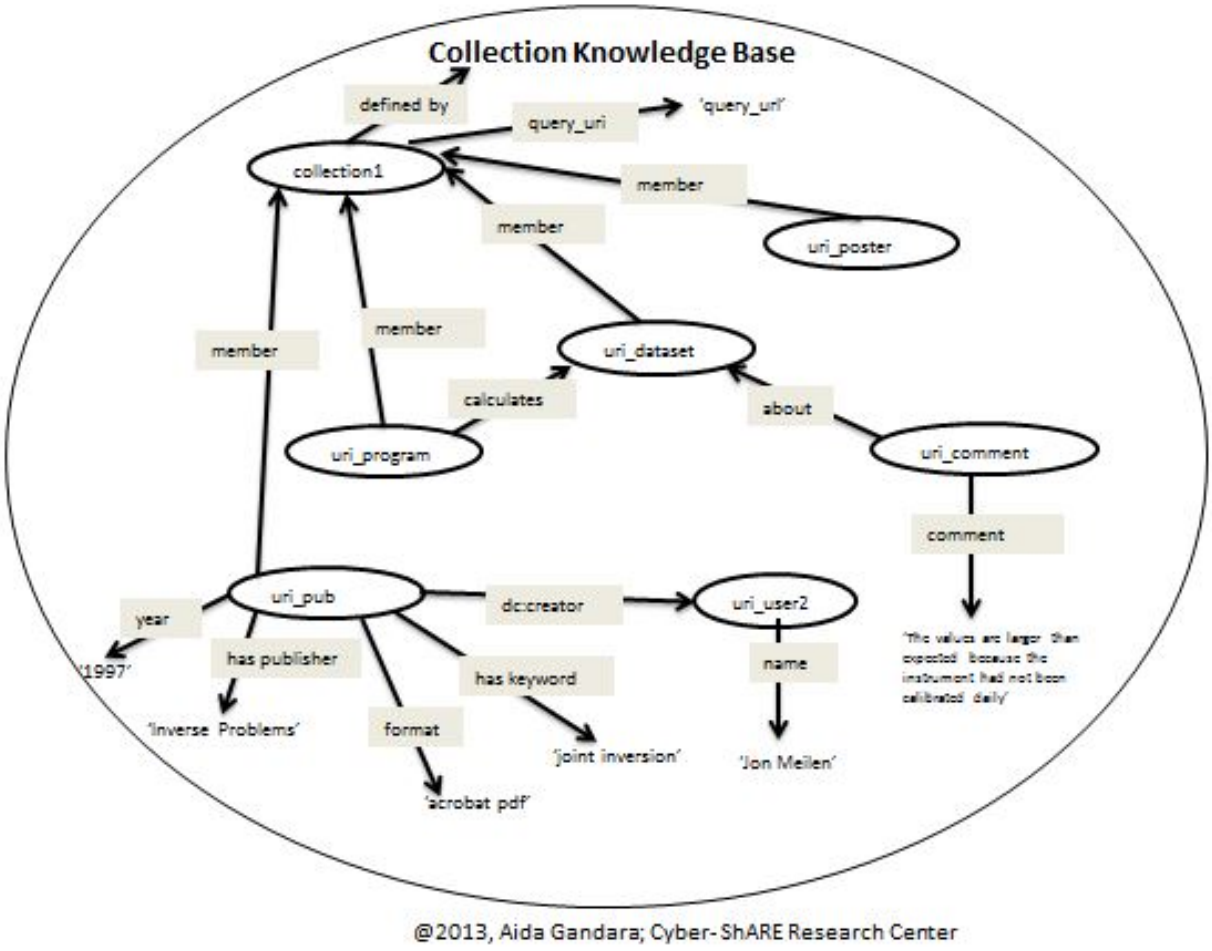

Figure 3.4: A graph representing a scientific collection that reflects Figure 3.1. In the ovals are the named individuals and the gray boxes represent roles. The arrows signify the relationship between objects or values. Not all properties or individuals are shown, for readability.

Chapter 4

Implementation of CARP Methodology

In order to validate and verify that CARP can, in fact, describe a scientific research effort, a prototype system was implemented. The prototype is an enhanced version of an existing Content Management System (CMS) that, by default, does not support several of the features that are needed for CARP. This chapter introduces the system and how it implements CARP.

4.1 CARP Prototype

The CI-Server Framework (Gándara, 2012), the initial research that set the foundation for this Semantic Web-based Methodology, is the base platform for implementing CARP. CI-Server is a set of programs, called modules, written in PHP that run on the Drupal Content Management System (CMS) Framework. Drupal supports customizations in content management, menus, user privacy, and other server details. Drupal is coded in PHP and it is open source (Drupal, 2013). The customizations implemented in the CI-Server Framework align with requirements to support CARP, such as: managing lists of Web-accessible resources called projects on the CI-Server, customizable project views through URLs and a client API so tools can be instrumented to interact with a CI-Server. The following sections describe the enhancements to the initial CI-Server Framework to support CARP.

Drupal has an estimated 900,000 user base[1] and over 20,000 registered modules. Although there are existing, albeit few, content management systems that manage semantic informationLausen et al. (2005); PublishMyData (2013); Kraker et al. (2011), several factors characterize why those systems were not selected as the CARP prototype platform. For example, some provide limited information and user management options, some are not open source and some are prototype systems. This research considers one approach to applying semantic information management into a commonly used environment and the changes that make the environment compatible with CARP. Such a transformation can occur in other types of systems as well, not just Drupal.

Drupal Content Management System

In Drupal, nodes identify content to be managed on the Drupal server. Nodes are configured per Drupal server as content types with associated attributes. Attributes can be of different types, including integers, strings, URLs, files and more. Modules installed on Drupal can add additional support for attribute types. Nodes are, therefore, instances of a Drupal content type. Drupal's support for RDF is implemented as modules that map node URLs, attributes and their respective values, to RDF triples, i.e., they map each node to subject, property, object triples based on attribute settings that are configurable to existing vocabulary. The mappings are imported from published vocabulary on the Web and the mappings are configurable. The Drupal RDF implementation uses the ARC2 triplestore to manage the RDF triples, using one triplestore for all Drupal node mappings. Thus nodes are the resources of the RDF triplestore. The existing Drupal RDF implementation was not utilized for a few reasons. First, if resources are self-described on the Semantic Web, the RDF implementation does not access or reuse that information. Second, if node file attachments have semantic information, the RDF implementation does not load this. Third, if attributes of the node have relationships (hyperlinks) to other resources on the Web, the RDF implementation does not use this information. Finally, the RDF implementation

[1] https://drupal.org/project/usage/drupal

has no support for grouping triples, as would be needed to distinguish different scientific collections.

The CI-Server implementation uses nodes in Drupal as resources in scientific collections. URIs can be set such that the node attributes are mapped to RDF to describe resources already on the Web, allowing node attributes to simply supplement the attributes of the resource. In addition, the CI-Server implementation enables the content of node file attachments to be loaded as RDF for a scientific collection, extracts hyperlinks found in node attributes to explicitly create **seeAlso** links between nodes and the hyperlink reference and groups the triples related to a scientific collection so that the collection is managed separate from other scientific collections. In addition, the CI-Server implementation can evaluate the URI setting and load triples from the source. Enabling CI-Server to load embedded RDF, self-describing URIs and SPARQL endpoint.

Figure 4.1 shows a table with the fields and values of a Drupal node as well as a graph that results from mapping the table to RDF. Drupal nodes store information to create and update resources as members of a scientific collection. In CI-Server, nodes have a URI attribute, if the attribute is set then the URI provides the Web location to access and harvest semantic descriptions for resources. If the URI attribute is not set, then the node attributes are the only properties captured for the resource; this is how resources that do not have a Web presence are introduced into a scientific collection. If later, a URI is identified, for example, the resource is located on the Web, then the URI attribute can be set in the Drupal node and the resource in the scientific collection is updated.

Multiple triplestore architecture

To reference specific URIs while building the scientific collection, the triples of the scientific collection are managed in a triple management system, e.g., a triplestore. ARC2 (Arc2, 2013) is a lightweight triple management system that supplies an API for managing triplestores, querying triples and updating triples, in addition to other Semantic Web-based capabilities. ARC2 is implemented in PHP and enables triplestores to be created program-

Drupal Node

Property	Value
URI	http://scientificdata.org/dataset01
description	'a dataset about the migration of birds'
creator	http://collabserver.org/person/Aida_Gandara
species	getSpecies()

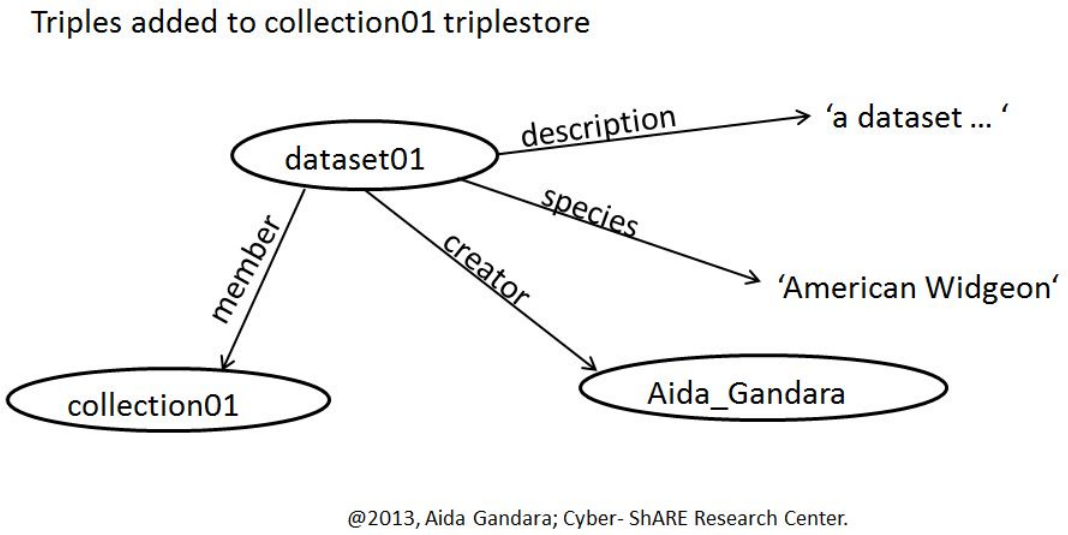

Figure 4.1: Sample fields of a Drupal node to an RDF graph

matically. By default, Drupal runs over a system-wide relational database to manage Web content. There are many modules on Drupal that work with the Drupal database, not triples or ontologies. CI-Server does not replace the relational database, the database is supplemented with triplestores. Basically, tools that run over CI-Server can still use the system and content in the relational database, however, they would need to be enhanced to take advantage of the semantic information in triplestores. The triplestores store Drupal information mapped into triples and they manage triples harvested from the Web. The ARC2 triple management system is integrated into Drupal and an ARC2 triplestore is op-

tionally created, one per CI-Server project. Figure 4.2 shows a diagram of the multiple triplestore architecture designed for the enhanced CI-Server.

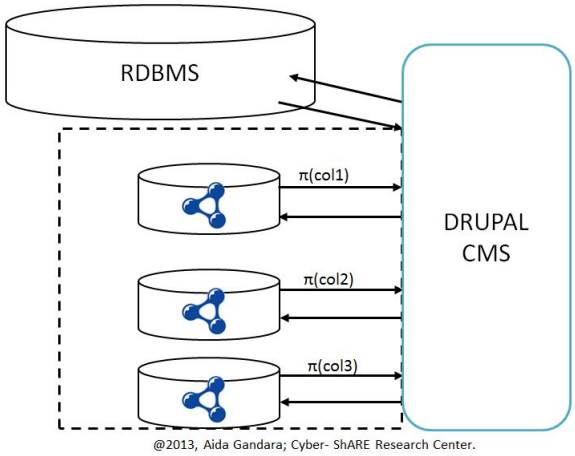

Figure 4.2: The CI-Server multiple triplestore architecture. Each scientific collection is held in a triplestore that supplements the Drupal database. Scientific collections can be accessed and queried individually through a SPARQL-endpoint.

For CARP, using a triple management system enables resource URIs to be explored, facilitating the ability to specify explicit relationships between them. Each triplestore has information specific to a project, the information is not shared across triplestores unless it is part of multiple projects. Changes to a scientific collection only affects the project-level triplestore. Separating scientific collections per triplestore is useful if one project loads triples that create a contradiction or have syntactic issues that render the triplestore unstable. In a system-wide triplestore such issues affect all triples, in the case of the CI-Server implementation, such issues only affect one scientific collection. For a CI-Server project, a triplestore holds a scientific collection that is exposed as a separate URI with a SPARQL query interface (SPARQL endpoint). CI-Server controls access to each SPARQL endpoint by exposing them through the Drupal menu system and the Drupal menu system adheres to Drupal user permissions. In addition, the triples of the scientific collection are accessible through a URI, i.e., exposed as a self-describing URI.

Core CARP vocabulary

To provide similarity across scientific collections, a core ontology is published on the Web to describe relationships and other common characteristics. All scientific collections can, by default, be queried using this vocabulary.

The CARP vocabulary is available to the CI-Server system and is used in the CI-Server modules[2]. Figure 4.3 shows a graph of the ontology. The graph shows a few classes with arrows connecting them; the arrows signify relationships between classes. The labels over the arrows identify the relationships. The class name is identified in the node with a prefix for the namespace. For example, three classes within the diagram are introduced in the CARP namespace as the following terms **ResourceCollection**, **DataServer** and **Annotation**.

Other vocabularies were considered. For example the Description of a Project (DOAP) RDF schema for describing software projects, such as open source software projects (Dumbill, 2004), the Simple Knowledge Organization System (SKOS) RDF schema that introduces common vocabulary for sharing and linking knowledge organization systems via the Semantic Web (Miles and Bechhofer, 2004) and the Vocabulary of Interlinked Datasets (voiD) RDF Schema for describing linked datasets (Cyganiak et al., 2011). Had either been chosen, scientific collections could assure integration and processing to similar data and tools. For example, there are tools that can analyze links within datasets that use the voiD vocabulary. As will be seen in 5.1, a bottom up approach was used when documenting research, meaning that vocabulary was added if it was needed to describe scientific data, not as a structure dictated for all scientific collections (Hausenblas, 2011). Since scientific collections can evolve as information is added, those vocabularies can be added if needed or requested.

[2]The CARP ontology can be found at: http://cybershare.utep.edu/ontology/2012/01/carp-ns.rdf

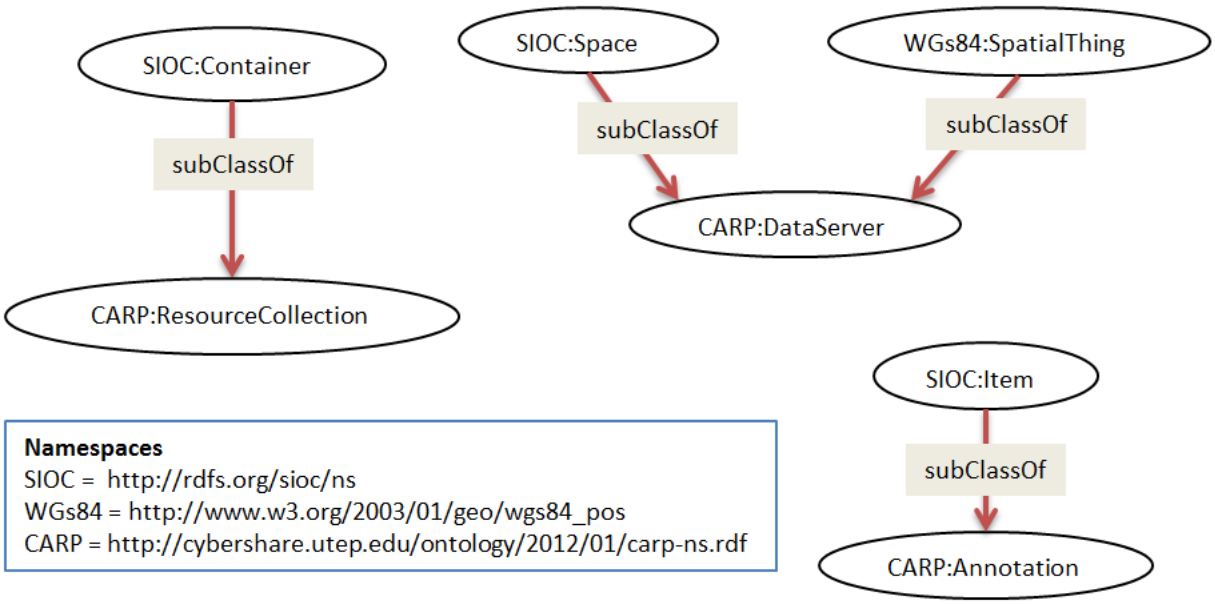

Figure 4.3: The core CARP vocabulary defines classes and properties used by the CI-Server implementation of CARP. CARP concepts inherit from other ontologies; respective namespaces are listed in the Namespaces box. The arrows between each class represent relationships, identified by the labels.

CARP Modules

As mentioned in Chapter 3.3.1, CARP is the result of existing efforts to document scientific research. The initial implementation is discussed in Appendix A. To enhance the initial version of CI-Server that did not integrate with the Linked Data dataspace, CI-Server modules were reorganized into five modules to support the different phases of CARP. When code is added to support a phase, it is added to its respective module. Moreover, a new module can replace some or all of the functionality of the the current CARP modules. The CARP modules are:

- Collect Module: This module supports the creation of resources to be added to scientific collections. The URI field of a node determines if a resource is Web-accessible or

not. If there is a URI setting then ARC2 tools are used to extract self-descriptions or embedded RDF from the resource URI. The additional fields on the node enable additional attributes to be set for a resource, for example fields such as creator, location, date information and a description. The fields are automatically mapped to Dublin Core, WGs84, etc. vocabularies. An attribute setting can be a function, in which case the function is applied to the resource file to extract a value. The functions must be PHP functions, registered in the CI-Server modules.

- Annotate Module: This module consists of two main functions. The first extracts hyperlinks from properties of the resources in a collection. The hyperlinks create links between a resource and the URL of the hyperlink. The second function obtains all comments for resources of a collection, creates a **CARP:Annotation** object and sets attributes for contributor, timestamp, about and previous. This module utilizes the default comment behavior of Drupal to capture comments about resources.

- Refine Module: This module captures relationships between two URIs and enters them into a scientific collection. Properties for relationships can be selected properties that exist in a scientific collection or they can be added to the local collection namespace. In addition, this module supports loading RDF from a URI into a scientific collection. The list of relationships and URI loads are stored in a file. ARC2 tools are used to extract embedded RDF from URIs, however, additional code can be added to the module to extract RDF from a URI using an RDFizer from the Semantic Web community or a custom PHP RDFizer function created in the **R**efine module.

- Publish Module: For each scientific collection, this module provides a URL for a SPARQL endpoint, self-describing URI of the scientific collection and an HTML page listing details about the collection. Additional functions can be added to the menu identifying the URLs that accept queries or return an encoded result set.

- Semantics Module: This module supports the underlying work of creating and synchronizing scientific collections based on what is created through the CARP modules. Each module interactively updates a scientific collection and also stores sufficient information such that it can be recreated and updated with existing Web content. For example, a Drupal node has the field mappings and URI of a source node and the Refine phase has a list of relationships and URI loads in a file. The Semantics Module initializes the base CARP menu, provides PHP calls to manipulate scientific collections, e.g., to add resources, relationships and properties, and implements the default mappings between nodes in a CI-Server project to triples in a collection. By default, the vocabularies used to map resources are FOAF, Dublin Core, WGs84, SIOC and the base CARP ontology. If a property does not have a namespace, then the property is added to scientific collection namespace. In addition, Drupal tags can be applied to categorize nodes within the Drupal system, these are mapped to the **Dublin Core:subject** relationship with the value being the tag string.

Shared repositories

To support common resources that have no self-descriptions on the Semantic Web, shared repositories, i.e., triplestores, have been created on the CI-Server. The common repositories cache semantic resource information and expose it through a SPARQL endpoint enabling software agents to query for it. This approach is meant to faciliate the harvest of semantic information. For example, a person's name, can be searched for in a shared repository containing information about people, by searching for **FOAF:person**. If found, a result Web-accessible URI is returned. The URI can be accessed for more information about the person or properties about the person can be accessed from the triplestore.

CARP in Drupal

Figure 4.4 shows the CI-Server implementation. Nodes are added to projects through the Collect Module. Comments and annotations are added as triples in the scientific collec-

tion through the Annotation Module. Additional resources can be added to the scientific collection through the Refine Module. The scientific collection is exposed as a SPARQL endpoint, a self-describing URI and encodings available through the URLs identified through the Publish Module. The Semantics Module orchestrates the creation of scientific collections by leveraging the different modules to generate triples in an ARC2 triplestore, i.e., by converting the nodes in the projects, extracting comments and annotations and loading refinements.

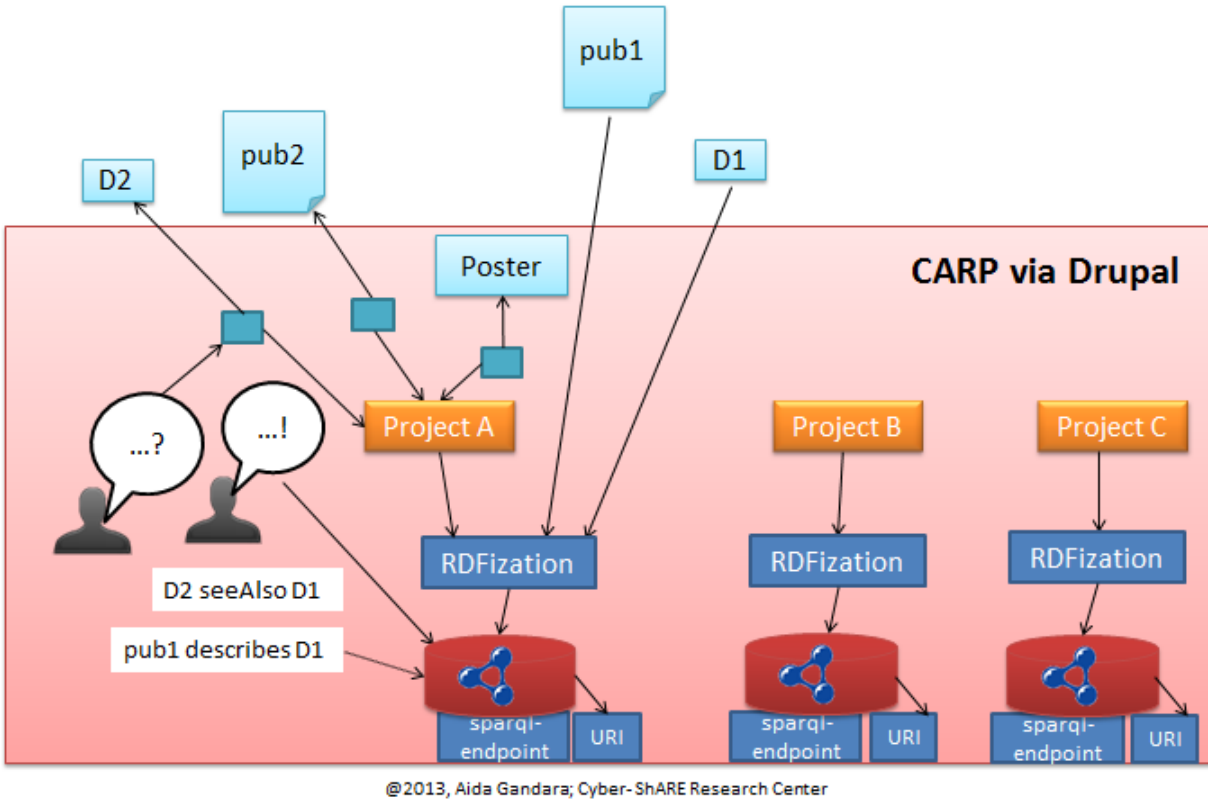

Figure 4.4: The CI-Server Framework on the Drupal CMS. Nodes and web-accessible URIs are RDFized and stored in an ARC2 triplestore, representing a scientific collection. Comments, annotations and other refinements are also added as triples into the collection. A SPARQL-endpoint and self-describing URI are exposed on the Semantic Web.

4.2 Related Approaches to Support CARP

This section describes six approaches that publish scientific research results on the Web and Semantic Web. Each approach is considered based on publications and information from respective Websites. The descriptions of each approach conclude with a few examples of how CARP can integrate with how those tools share on the Web or Semantic Web.

For each approach, the following characteristics are considered:

- How the approach collects resources
  1) Does the approach build a collection of resources?
  2) Does the approach harvest information from the Web?
  3) Is information semantically structured, i.e., through an ontology?
  4) Is the URI preserved for referenced resources?

- How the approach adds annotations
  1) Are hyperlinks identified from resource attributes?
  2) Are comments captured and specifically related to resources?

- How the approach specifies relationships
  1) Can relationships be added with resources in the local system?
  2) Can relationships be added with URIs on the Web?
  3) Is the vocabulary flexible, i.e., can any vocabulary be used to specify a relationship?

- How the approach exposes information on the Web
  1) Are original Web URIs published?
  2) Can the information be accessed or queried by machines?
  3) Is semantically structured information exposed?
  4) Can the information be exposed in different formats or encodings?

Table 4.1 provides a summary of the approaches. The research proposes that the information shared on the Web or Semantic Web by the tools listed in Table 4.1 can be reused

and enhanced by scientific collections created by CARP. Given that scientists share their research to make it accessible to others and to facilitate reuse, enabling use of mechanisms that support information integration, data exchange and understanding of scientific results should increase relevance of the Semantic Web to scientific research and encourage participation in the Linked Data dataspace by scientific data providers and consumers. In addition, most of the approaches mentioned can implement some or all of the phases of CARP to enhance the understanding of scientific resources as linked data.

Data portals

Data portals capture relevant metadata about scientific data and make resources accessible on the Web. Most data portals have software infrastructure to capture datasets from scientists and provide search tools to support locating and downloading scientific datasets. One such example is the Knowledge Network for Biocomplexity (KNB), a national network of federated institutions that share data and metadata using a common framework that works around the Ecological Metadata Language (EML) for describing ecological data and the Metacat metadata server for storing and sharing XML metadata documents. KNB consists of many sites that participate in the KNB network, for example, sites from the Organization of Biological Field Stations (OBFS), another collaboration between institutions. The infrastructure provided by KNB provides search, data integration, quality assurance and visualization tools for published datasets. In addition to the collaborations within the KNB, recently, KNB has extended offering data holdings to DataONE, another group of organizations that are unifying efforts to offer earth-based datasets. The DataONE collaboration has established member nodes to share data and server nodes to manage the member nodes and a DataONE API to automate the interaction between the different nodes.

Other organizations such as the ADIwg have similar intentions to increase integration and exchange across collaborating members. Each effort identifies some protocol and vocabulary for exchanging information across organizations, and this could change as new collaborations are considered. Protocols for accessing resources can range from openly

available on the Web to password protected API calls or services, and most of the vocabularies used to exchange information are XML-based. In addition, each collaboration handles accessing and searching for resources appropriate for the domain of data. For example, the ORNL-DAAC enables searches for data using a map-based search interface called Mercury. Other functionality such as annotating or create links between datasets or to external Web resources is not common practice. Neither is the ability to expose datasets in different serialized formats, e.g., from EML format to a different metadata format compatible with the ADIwg group.

Data portals do not usually support creating collections of datasets, harvesting information from the Web or managing semantically structured information. Most information is managed in a relational database. If there relationships to other resources are captured, it is normally through an attribute that that references a Web document through a hyperlink, not through URIs, however, the data portal system does not make use of the hyperlinks. In addition, comments are not normally added to datasets, aside from comments that might be captured in the metadata of a dataset. Data portals define the metadata that is added with a dataset, in order to support consistent searches and views of the metadata in the data portal system. Generally, data portals enable some type of automated search, either through an API or services. The information exposed for a dataset, either through a Web page or an API is generally not semantically structured, although some data portals do support semi-structured views, e.g. an XML formatted URL encoding. Aside from an HTML view or API calls, there is generally no support for exposing metadata or datasets in additional encodings.

CARP can function to extend the information provided by a data portal. First by harvesting RDF about a dataset. If the description of the dataset is in XML or RDF, such an extraction is easier, otherwise, tools can be created to consistently extract information from a data portal. For KNB, for example, datasets can be RDFized by mapping the EML fields to RDF vocabulary, such a transformation has been considered for EML[3].

[3]A study from Syracuse University, The EML project, provides an RDF faceted search and RDF files

Once the dataset is in the scientific collection, the remainder of the CARP process can be followed, enabling users to identify URIs and allowing for annotations and relationships to be captured about them. Future mappings between the information in a dataset to other data providers or data standards can focus on vocabulary mappings between the scientific collection and other vocabulary, without having to choose new data structures or APIs.

Figure 4.5 shows the scenario where information from data portals are RDFized into a scientific collection. If a data portal were to provide the structured description, i.e., in RDF, the portal can still require logins to access information for integration and exchange or the portal can expose a structured and non-secure version to enable general browsing of dataset metadata, requiring passwords when downloading data. Many datasets can be added to a scientific collection, as can other research objects, making related datasets and other research objects available and searchable from a single scientific collection regardless of the data provider.

Research Objects

Research objects(ROs) are collections aggregating multiple resources that result from scientific research. ROs are stored as archives (files) consisting of research resources collected from a directory or referencing URIs on the Web. Resources collected in the RO are identified with non-information URIs. An RO template defines the expected structure of the archive as well as the vocabulary for semantically describing the RO in RDF. Different types of research objects are described with predefined RDF-based vocabulary, depending on their purpose, i.e., scientific social objects (De Roure et al., 2011) and workflow-centric research objects (Belhajjame et al., 2012). In particular, the focus of ROs is to document resources related to workflows and upload them to a Web server for others to access and, in some cases, execute. Annotations can be added through controlled vocabularies that represent tags, links to text and links to URIs. ROs can be uploaded to the myExperiment environment (De Roure et al., 2007) and queried through a system-wide SPARQL

created from EML files. The results are at: http://sdl.syr.edu/eml/index.html

endpoint. Workflow-centric ROs can be managed through a workflow preservation infrastructure called the Wf4Ever Architecture, meant to support search, storage, management and analysis, workflow lifecycle and more.

ROs do not harvest information from the Web, the information in an RO is described with a semantic vocabulary and URIs referenced in a RO are preserved in the RO definition. Comments can be added to ROs; comments appear to be added to an RO generically and not related to a resource in the RO directly. Hyperlinks that may be referenced in a comment are not added as relationships to the RO. Relationships can be added to an RO between a resource and other internal or external URIs, using a predefined vocabulary for the type of RO. ROs can be queried and their vocabulary exposed if they are loaded onto the myExperiment environment that shares the ROs in a SPARQL endpoint. ROs are not exposed in different formats or encodings. ROs are capable of participating in the Linked Data dataspace because they are in RDF, although, the environments that share them do not facilitate capturing annotations or relationships between URIs or enhancing ROs with additional vocabulary that can be used to exchange or integrate ROs with other Web resources.

CARP could function to extend the information provided for a RO. This is shown in the capture of RDF from ROs to the scientific collection in Figure 4.5. Since ROs are already structured in RDF, their URIs and related RDF can be imported into a scientific collection. The URIs can be used in annotations and additional relationships can be applied, for example, to use different vocabulary not allowed by the template of an RO and by assuring that comments relate directly to resources in the RO. Performing these operations on an RO enables an RO to be accessible from the collection, along with other research resources that are related to the RO.

Semantic Web Servers

Semantic Web servers manage information as triples and some expose that information on the Semantic Web. Some of these systems manage semantic information in order to facili-

tate gathering information but do not expose semantic information on the Semantic Web, others only expose SPARQL endpoint and others support Linked Data principles, where semantic descriptions of resources can be accessed through self-describing URIs. RDF Extensions (RDFx) (Corlosquet et al., 2009), formerly RDF CCK, is a Drupal implementation that embeds RDF support into the Drupal Content Management System using various modules(Corlosquet et al., 2009). The main purpose of RDFx is to expose the structure of a Drupal site to the Semantic Web, in particular exposing Drupal nodes and comments as linked data. RDFx does not provide any specific support for collecting resources but does share selected nodes through a SPARQL endpoint. RDFx provides default RDF mappings for nodes but allows system administrators to customize the mappings to other ontologies. Identifying ontology mappings is supported by an external ontology search service and vocabulary importer service. As nodes are added to the system, the mappings are used to encode node Web pages with RDFa and optionally to create triples that are stored in a system-wide ARC2 triplestore.

There are different ways of configuring RDFx. Vocabulary changes, e.g., adding ontology terms that can be used in node mappings, are implemented through edits to an ontology and manual mappings of Drupal nodes. A site-based vocabulary is created on the Drupal server to describe the constraints on fields and types from the structured schema of the database. Annotations are supported through Drupal comments that relate to nodes and are mapped to the **SIOC** vocabulary. RDFx supports loading additional RDF into the triplestore through a SPARQL Proxy that runs queries to load RDF on demand. The SPARQL Proxy commands are configured by a system administrator; executing them loads triples into the triplestore and links URIs. The triplestore can be exposed as a SPARQL endpoint, enabling software agents to query the triplestore.

RDFx are currently used by approximately 799 Drupal sites[4]. Since Drupal is not, by default, built to support semantic information, RDFx is the default approach. However, there are limitations in what RDFx can support. For example, related information, e.g.,

[4]statistic taken from RDFx Module Drupal page at https://drupal.org/project/rdfx

files and hyperlinks, are not included in the RDF of a Drupal node. In addition, RDFx, or related modules, does not support adding links between resource URIs. In addition, RDFx does not support serializations to different formats.

CARP can access the RDF provided in RDFx to describe resources in a scientific collection. Then enable annotations and additional relationships to be added. Since RDFx already produces linked data, the implementation of CARP can leverage the RDF and SPARQL endpoint that support RDFx. This is seen in the capture of RDF from ROs to the scientific collection in Figure 4.5. The scientific collection would use the URIs provided by RDFx to add more research related details. Moreover, CARP can be implemented as part of RDFx. Since RDFx runs over a system-wide triplestore, scientific collections would still require their own triplestores or the CARP implementation would need to assure that queries into a system-wide triplestore are capable of producing results specific to individual scientific collections.

Linked Open Data producers

Some efforts to share scientific data are doing so as open data. Many of these have identified themselves on the linked open data cloud. One example, Bio2RDF, has converted heterogeneously formatted biological data (e.g. flat-files, tab-delimited files, SQL, dataset specific formats, XML etc.) into RDF and RDFS. Once converted, the biological data can queried through a SPARQL endpoint or downloaded as RDF. The Bio2RDF project currently hosts over 1 million triples for 19 datasets[5]. The Bio2RDF system does not support collections of datasets, each dataset is captured separately, although adding more data to a Bio2RDF dataset could occur manually.

The Bio2RDF datasets are supported by a set of scripts that function to ensure a high level of syntactic interoperability between the generated linked datasets, e.g., by creating valid Bio2RDF resources and only making use of preferred namespaces in a dataset. In addition to following Linked Data principles, Bio2RDF adheres to the policies of the

[5]https://github.com/bio2rdf/bio2rdf-scripts/wiki

Banff Manifesto. Bio2RDF datasets are annotated with the W3C voiD vocabulary, the Provenance vocabulary (PROV) and Dublin Core vocabulary to document provenance of the datasets. The provenance captured enables datasets include attributes about dataset creation and other statistical values.

One premise for creating the Bio2RDF data stores is the need to integrate results with other related datasets from other organizations. Nevertheless, Bio2RDF infrastructure does not provide mechanisms to annotate or relate information in the datasets to other datasets. Moreover, there do not appear to be mechanisms to refine the datasets to support integration or mappings to other datasets or RDF vocabularies.

CARP can be used to enhance the datasets created by Linked Data dataset providers such as Bio2RDF. First by importing URIs from a dataset and then enabling the information to be enhanced with annotations and refinements. This is shown in Figure 4.5 where Bio2RDF RDF is imported into the scientific collection. Since the datasets are quite large, a CARP implementation could optionally select a smaller set of related information to import into a scientific collection. Accessing more information about a dataset would be a matter of accessing the self-describing URI from the scientific collection. In addition, the Bio2RDF datasets would be grouped with other resources in a scientific collection, further enhancing how datasets relate to other resources of a research effort. Finally, **P**ublish can support the encodings of the datasets to integrate with or map to other types of data or standards.

Linked Data Look-up Servers

Another technique considered in this section are document lookup systems. Some of these systems only support searches over the Web, for example, using sitemaps to search for documents. However, some work over structured data, using Semantic Web techniques to search for Web URIs. One example is Sindice, a tool created to help application developers in locating data sources and allowing them to connect sub-graphs of related data sources. Sindice manages a large collection of indexed URIs and keywords, loaded from RDF doc-

uments. Sindice provides an API for integrating information and SPARQL endpoint for accessing information.

Sindice provides a Webpage for searching a system-wide triplestore, analyzing the RDF graphs and requesting more dataset uploads. Sindice is a simple RDF lookup tool. As such, it does not allow for creating and analyzing collections of related resources, annotations of resources or refinements that might facilitate resource lookup. Such information would need to be loaded from an RDF document and only the URIs and keywords are accessible from the Sindice system. In addition, there is no way to configure the environment to support specific vocabulary or vocabulary mappings that might make the triples more useful to a specific context. For example, to map resources to a different vocabulary that would facilitate analysis using other tools.

CARP can enhance the indexing provided by a document lookup system. In this case, Sindice is already in RDF so this facilitates the **C**ollect process, where the lookup system can be used as a data source, i.e., the SPARQL endpoint can be searched and URIs as well as their properties can be optionally added to a scientific collection. Figure 4.5 represents this with the arrow leading from Sindice to the scientific collection. CARP can add more meaning about the indexed URIs. Adding annotations and refinements to describe what is not available in the loaded RDF documents would add meaning to the Linked Data dataspace. In fact, relevant information from a scientific collection could be loaded back into a Sindice dataset.

Social Tagging System

Environments for social tagging enable users to organize personal resource collections with tags (citeulike, 2013; Zotero, 2013; readcube, 2013). Social tagging systems can be used to collect Web resources of scientific significance such as images, datasets, and publications. The resource types are essentially unlimited, although specific social tagging system implementations limit collections to classes of resources, e.g., publications (citeulike, 2013) and images (Flickr, 2013). Social tagging systems import information by extracting resources

from Webpages or through users filling out forms, i.e., the URI is not usually preserved. Some social tagging systems enable comments to be entered about resources or as comments related to the entire collection of similarly tagged resources. Resources and collections of resources are available to users through the social tagging system interface and, in most cases, only available to machines through APIs that are specific to the system. Finally, there is limited use of ontologies and Semantic Web techniques, aside from using ontology terms to tag resources, and relationships between resources are not captured, aside from membership to a collection.

CARP can be used to enhance the resources identified in a social tagging system by importing the collection data, adding annotations and relationships and then exporting the content in a scientific collection. If the social tagging preserves the Web URI then CARP is able to enhance the Linked Data dataspace, otherwise the result scientific collection would have information that is not related to the Web. Since social tagging systems collect and manage the information added to a group of similarly tagged resources, social tagging systems can implement some or all of CARP to enhance knowledge about collections of Web resources. Social tagging systems could then exchange information with client tools that make use of their collection.

Table 4.1: The table considers how each approach collects, annotates, relates and publishes information on the Web or Semantic Web.

Approach Characteristics ====>	COLLECT 1) Build collection? 2) Harvest from Web? 3) Semantic structure? 4) Preserve URIs referenced?	ANNOTATE 1) Identify hyperlinks? 2) Capture comment to resource?	RELATE 1) Capture local relationships? 2) Capture remote relationships? 3) Allow flexible vocabulary?	PUBLISH 1) Publish original Web URI? 2) Access or query information? 3) Expose semantic structure? 4) Support different formats or encodings?
Data Portals	1) No 2) No 3) No 4) N/A	1) No 2) No	1) Generally, No 2) Generally, No 3) No	1) No 2) Generally, Yes 3) Generally, No 4) Generally, Yes
Research Objects	1) Yes 2) No 3) Yes 4) Yes	1) No 2) Yes	1) Yes 2) Yes, through URI 3) No	1) Yes 2) Yes 3) Yes 4) No
RDF Extensions (Semantic Web Servers)	1) No 2) Yes 3) Yes 4) Yes	1) No 2) Yes	1) Yes 2) Yes 3) No	1) Yes 2) Yes 3) Yes 4) No
Bio2RDF (Linked Open Data Producers)	1) No 2) No 3)Yes 4) N/A	1) No 2) No	1) No 2) No 3) N/A	1) Yes 2) Yes 3) Yes 4) No
Sindice (Linked Data Lookup)	1) No 2) Yes 3) Yes 4) Yes	1) No 2) No	1) No 2) No 3) N/A	1) Yes 2) Yes 3) No 4) No
Social Tagging Systems	1) Yes 2) Yes 3) No 4) No	1) No 2) No	1) No 2) No 3) n/a	1) No 2) No 3) No 4) No

4.3 Summary

All of the approaches mentioned above enable scientists to share research resources on the Web. Some even produce Linked Data. CARP suggests that enhancing information on the Linked Data dataspace will enable more understanding and therefore more reuse of related resources. For example, by using Semantic Web-based information integration techniques to compare two datasets or by locating similar resources based on their type or properties. CARP promotes extending the information shared by the tools described in this chapter to make related information become more accessible based on the context of the collection as well as uniformly understandable to machines.

Figure 4.5 shows the contribution of using CARP to describe scientific collections based on the approaches described in this section. Figure 4.5 shows both structured and unstructured information added to a scientific collection as RDF, enabling the meaning of resources that heterogeneous to be processed similarly.

Documenting distributed and heterogeneous resources in the context of a scientific collection is expected to have additional benefits from other areas of research. Three general examples are shown in Figure 4.5: 1) tools implemented in CARP can be used to map a scientific collection to other representations and exposed as a URL, i.e., from EML to constructs in the ADIwg metadata; 2) existing mappings can be shared as part of the scientific collection, facilitating integration and information exchange for different tools; and 3) Semantic Web-based tools can analyze the content of a scientific collection to facilitate query construction of scientific research, e.g., through natural language interfaces.

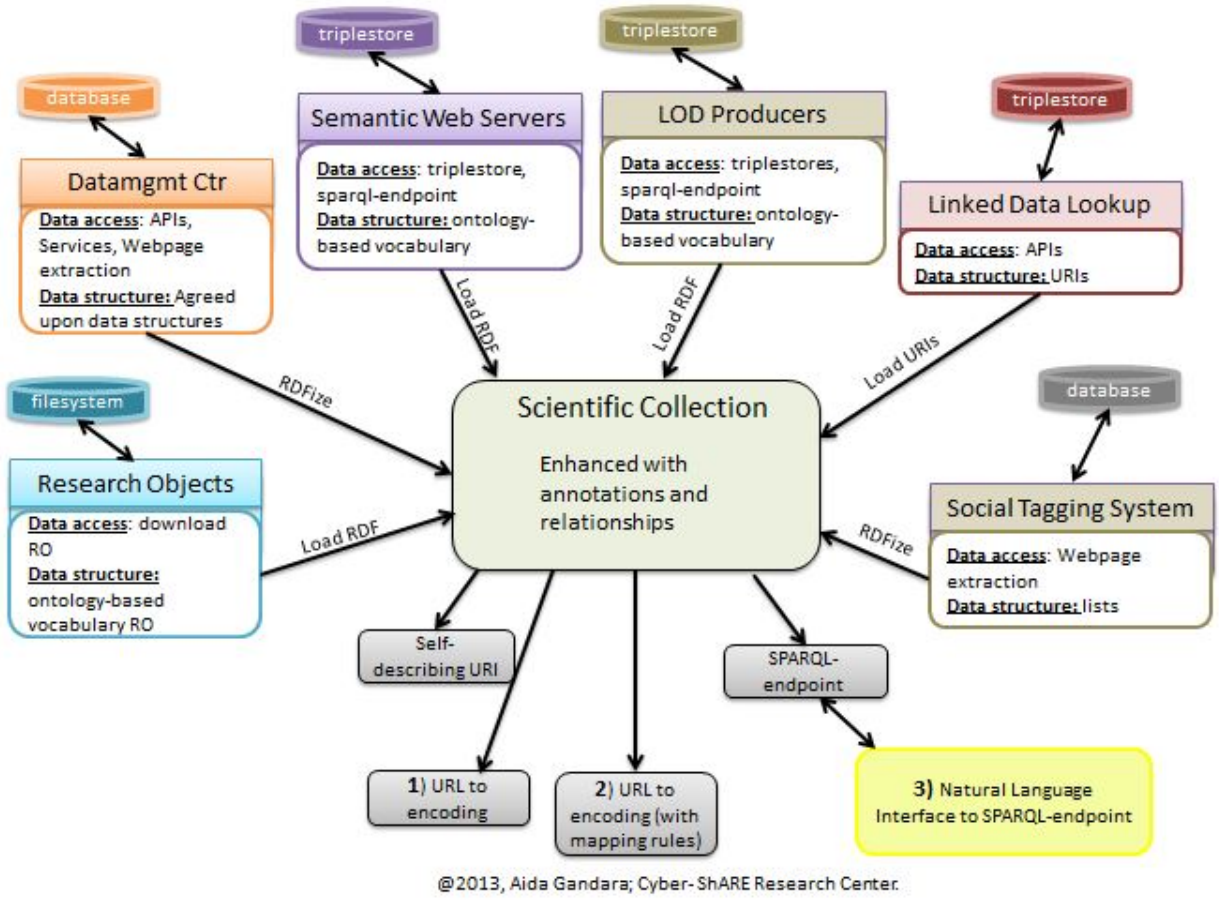

Figure 4.5: CARP reuses information on the Web making scientific collections accessible as Linked Data.

Chapter 5

Validation and Verification

Scientific collections are knowledge bases that describe details about scientific research. To assure that CARP can document scientific research and to get a better understanding as to how actual scientific teams could leverage CARP, three case studies were conducted and are documented in this chapter. Finally, the case studies discuss the verification of scientific collections through the application of research specific SPARQL queries and by checking the ontologies with an ontology validator. This research has introduced an approach to documenting scientific research as scientific collections that leverage the Web and Semantic Web to share descriptions about scientific research. CARP consumes information about resources on the Web to produce Linked Data. To do this, the methodology consistently adheres to the five CARP principles to **C**ollect resources out on the Web, identifying and adding URI specific **A**nnotations and **R**efinements and **P**ublishing the information back onto the Linked Data dataspace. This chapter elucidates how this approach enhances what is already on the Linked Data dataspace; ... and finally, to assess the opinions of Cyber-ShARE researchers on the benefits of scientific collections in increasing accessibility, understanding and attribution of scientific results, a survey was conducted and is presented.

5.1 Case Studies

The Cyber-ShARE Center has three primary research projects, mostly funded by the National Science Foundation (NSF) (Cyber-ShARE, 2012), that can benefit from a methodology to document research results. Research for the Center is ongoing between the principal investigators of the Center and smaller research efforts consisting of undergraduate and

graduate students from the University of Texas at El Paso. One issue for the Center is keeping track of the many results and findings of researchers after they complete their studies and leave the University. For example, there is minimal documentation about each research effort that accurately identifies all collaborators or the location of digital datasets that result from scientific studies. A small-scale survey with the members of two CyberShARE research projects identified the resource types and current storage locations of their research results, as well as their opinions about sharing their research. In addition to a general consensus by the surveyed group that most of their resources are not actually published on the Web, this survey found that most of the researchers agreed that accessible and understandable research resources are important for research reuse. The majority of the group also agreed that they needed to increase accessibility and understanding of their research results. Details of this survey are found in Appendix B. To understand how CARP applies to research efforts at the Center, three case studies were identified to systematically step through documenting their research results.

Each case study was conducted with the goal of answering the following questions and sub-questions:

1. What resources do researchers capture that could document their research?

   (a) Is there attribution in the resources?

   (b) How many of these resources have semantic descriptions?

   (c) How much collaboration can be identified from resources that are publicly shared?

   Questions 1 and 1(a)-(c) consider the state of the resources that are relevant to documenting a scientific research effort. The initial survey contains the list of resource types that are captured by researchers at the Center. Question 1 confirms the list of resources for a case study and the sub-questions confirm additional qualities of the resources. Question 1(a) considers where the authors or owners of a resource are explicitly identified in the resources. Tools for data capture and data curation are

increasingly adding structure to data. Question 1(b) considers if the structure that potentially exists in the resources of a case study have semantic structure, i.e., are they structured using ontologies for use on the Semantic Web. Finally, to support the Center in identifying collaboration within research efforts, Question 1(c) considers whether those resources that are Web accessible account for the different collaborators that researchers identify.

2. Are the phases in the methodology sufficient for documenting each case study?

   (a) Are the CARP phases followed to document each case study?

   (b) Does a case study have documentation needs outside of the CARP phases?

   Questions 2, 2(a) and 2(b) consider whether the methodology is effective in each of the case studies. The following two sub-questions capture more details to answering Question 2. If the phases are followed, Question 2(a) will be true and following CARP to document research resources will be straightforward. Since CARP is a high-level methodology, the expectations are that a resource can be collected and iteratively annotated, refined and then published as part of a scientific collection. Nevertheless, this question will be documented to assure confirmation of Question 2(a). Question 2(b) asks the alternate question, even if the phases are followed, are there documentation needs that the methodology does not support? For example, the methodology does not specify support for documenting collaborative research communities. If a case study has a documentation need the methodology can not support then answering this question should capture this deficiency.

3. What semantic vocabulary is used in the scientific collection for each case study?

   (a) Is it acceptable to focus on vocabulary at resource level instead of a predefined vocabulary?

   (b) Will scientific collections use default vocabulary or will additional vocabularies be needed?

(c) Will researchers request any specific vocabulary?

Question 3 and 3(a)-(c) consider the ontologies used to describe scientific collections. What is of interest is determining where the vocabulary should come from, a predefined and default vocabulary or will there be a need to accommodate the data and requests of researchers. Question 3(a) considers whether it will be acceptable to create scientific collections as bottom up, i.e., adding the vocabulary of resources in lieu of a fixed vocabulary predefined for all resources of scientific collections. Question 3(b) considers whether the default vocabularies provided for different resource types will be sufficient or will there be a need to add additional vocabulary for a case study. Question 3(c) considers if the researchers have specific vocabulary needs, e.g., if they need to integrate with a known community or if they have a specific request because of a standard they are interested in following.

Researchers of the three case studies discussed research and provided resources relevant to conducting scientific research. For each case study, the effort to work directly with researchers was limited to 3 to 5 hours, averaging about 3 total meetings per group and a few emails to clarify details. Figure 5.1 shows a diagram highlighting the progress of each meeting. The first meeting was to gain an overall understanding and perspective of the research. The primary researchers of each group were interviewed to discuss their work, its importance, the research goals and research questions. The intermediate meetings focused on obtaining clarification about resources, understanding relationships and obtaining data or other relevant resources that were not originally provided. The final meeting was used to show research team members a Webpage with all the research information collected, to show and discuss answers to questions and to obtain their input on the benefits they saw in sharing their research as a single scientific collection. CARP was leveraged to organize capturing the details about resources, e.g., attributes, annotations and relationships, and to identify code changes that would help automate, or semi-automate, the different phases. Throughout the case studies, notes were maintained on meetings and discussions and the

scientific collections captured details about resources, vocabularies and the research documentation in general. It should be noted that the researchers did not create the scientific collections primarily due to time constraints. Their neither had the time to review all the resources to conduct the case studies nor did they have time to learn a new system, prototype system, while they were conducting their research. The role of this research was in understanding their work while following and enhancing CARP using their research results, answers and discussions.

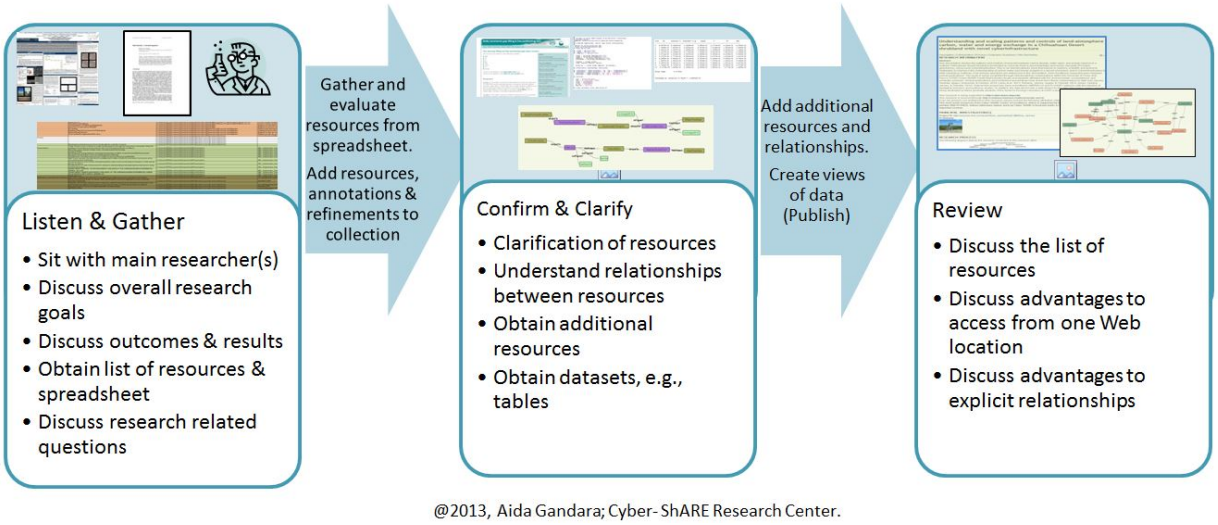

Figure 5.1: Summary of meetings held with the different case studies.

The following sections describe the case studies. Each case study section includes an overview, a description of the scientific collection, a discussion of how scientific collections are accessed, used and verified, and a summary about results.

5.1.1 Eddy Covariance Cyber-infrastructure (ECC) Project

Over the last century arid and semiarid regions have been affected by desertification, also known as shrub encroachment. Ongoing studies have been conducted to understand how land cover change impacts microclimate. Due to the poor understanding of the impact of these transitions, an improved understanding of ecosystem dynamics and land-atmosphere

interactions at local scales and capacity is needed to extrapolate the dynamics to regional scales using remote sensing. Eddy covariance (EC) systems could help in identifying the appropriate scales for accurately representing the footprint of eddy covariance flux measurements that would appropriately extrapolate, however, these systems are not yet well understood. Researchers from the ECC project are working to understand the factors for controlling land-atmosphere carbon, water, and energy exchange in a desert shrubland using standards and protocols developed by the eddy covariance and optical remote sensing communities and new Cyber-infrastructure tools adapted from other scientific fields. For three years, ECC scientists have conducted research and documented the process from the initial site selection phase and eddy covariance system implementation to eddy covariance data processing and visualization (Jaimes, 2012)[1].

Overview of ECC Project

The ECC project has, thus far, captured over 1 Terabyte in data and other files. The overall process for this research has produced large amounts of raw and processed data that could be valuable to future scientific research. ECC scientists have tediously stepped through the acquisition and processing of scientific data in order to document and compare results from differing eddy covariance infrastructures. As a result, researchers have captured detailed documentation of eddy covariance solutions covering a span of three years, 2010, 2011 and 2012. The details of the cyber-infrastructure and related studies are only partially described in publications and posters. A workflow is available as an image and provides a previous version of the cyber-infrastructure that is currently under consideration for the research effort. This project is ongoing which introduces challenges in how research results are managed and shared, for example, not all resources can be shared publicly while others can, in addition, different phases of the project are planned and others are active.

To discuss the ECC project and results, the ECC principal investigator (PI) was the

[1]The previous paragraph is a summary of the motivation behind the ECC project, collected from resources of the case study.

main contact, providing a spreadsheet of resources, research related questions and an organized directory structure with relevant research resources. The ECC project depends on different tools and programs, some created by the ECC project and others reused, to create the eddy covariance cyber-infrastructure. Some resources were excluded if they contained private information such as user names and password, if a resource was duplicate or if the research PI identified documents as no longer relevant or for internal use only. A list of collaborators was collected from the PI and from manually scanning resources. Resources were evaluated for semantic structure. There were no researcher initiated discussions about semantic vocabularies, aside from an initial introduction discussing the goals of scientific collections. However, there is research specific vocabulary to describe the phases of the research and cyber-infrastructure. There are plans by the ECC project to share datasets with other Cyber-ShARE researchers and other eddy covariance communities, e.g., FluxNet[2]. Initially, there was consideration for using research questions from the ECC project's proposal research questions, however, as this is an ongoing research effort, data was lacking to answer those questions. Moreover, the ECC project PI was particularly interested in answering questions about the eddy covariance cyber-infrastructure under study, e.g., the tools that are used and collaborators that helped build it.

ECC Project Resources

After an initial evaluation of all the files and resources provided by the ECC project PI, 238 resources were considered for documenting this research effort. Table 5.1 shows two tables, the top shows a summary of where 238 resources from this case study were located at the beginning of the documentation; the bottom table shows collaborator information captured from the different resources and discussions with researchers, showing collaborator qualities at the beginning of the documentation. 93% of the resources were located on *local and personal drives* which included local directories on a researcher's laptop, directories

[2]FluxNet is a component of NASA's ORNL DAAC that maintains a central database of fluxnet data, tower and site characteristics

on an external hard drive connected to a researcher's laptop and a files on a drive on SugarSync[3]. This count only includes files that were owned by the researcher. To make these accessible to document research, these were uploaded to a server and assigned a URI and added to the semantically described collection. Uploading these resources also assured that files were moved to Cyber-ShARE drives and not left on local drives. There were resources identified as part of the research effort but having no digital representation, these were identified as *Neither* in Table 5.1. For example, the raw sensor data collected at the eddy covariance site is large and not available to share on the Web, however, it is part of the eddy covariance infrastructure. Nodes were created for these resources and assigned a URI and added to the semantically described collection. *Neither* resources totaled 1.26% of all resources. *Online* resources included data that was generated by another tool, online workflows, publications available at a publication site and tools used to conduct research. For each, the URL for the resource was located on the Web and added to the semantically described collection. These totaled 6.3% of the resources identified for this research effort.

Attribution is of particular interest to the ECC project PI since the infrastructure is gathered from various sources. For each resource documented, the creator and owners were identified. This was a challenge, as seen in the 15.55% *attribution* average shown in Table 5.1. Posters and presentations were created for sharing, for the most part, and have a high attribution (96%), although there was one presentation that had no attribution. Images, animations and workflows were also created for sharing, however, researchers did not capture attribution for most of these. No attribution was captured for data. Publications were not always published for peer reviewed purposes, for example, one document provides step-by-step instructions on changing and configuring a sensor card at the Jornada Basin site. Consequently, there were documents with no attribution resulting in a lower attribution rate than for publications (77.78%). The ECC project utilized online tools and those were all considered as attributed since the creators were accessible through the URL,

[3]http://sugarsync.com

Table 5.1: The two tables summarize the state of resources at the beginning of documenting the ECC project case study. The top table shows the different resources, where they were located, their attribution and self-description. The bottom table shows the collaborator information calculated from the resources, including self-description, the collaborations found in some online document and participation loss.

RESOURCE TYPE	Local/ Personal	Online	neither	TOTAL	attribution	%attribution	self description
image /animation	17	0	0	17	3	17.65%	0
data	172	3	3	178	0	0.00%	0
workflow	1	2	0	3	0	0.00%	2
program	4	0	0	4	3	75.00%	0
tool	1	7	0	7	6	85.71%	0
presentation	14	0	0	14	13	92.86%	0
poster	5	0	0	5	5	100.00%	0
publication	6	3	0	9	7	77.78%	0
TOTAL	220	15	3	238	37	15.55%	2
%	92.44%	6.30%	1.26%				

COLLABORATORS	TOTAL	self description	found online	participation loss
people	21	2	13	38.10%
organizations	5	0	1	80.00%
TOTAL	26	2	14	

Note: this is the status before published on a Web server and added to a collection

however, one tool was provided by another student and had no attribution. Only 3 of the 4 programs had a name referencing the creator. Attributions were accepted loosely for programs, for example 'Creator: Jane' was an acceptable attribution if the creator could be understood by researchers. This was allowed since the focus was on a specific research effort and it simple to resolve the person's name to their self-describing URI in the CI-Server person repository.

Only two resources had any semantic representation because they are OWL documents that were generated by a workflow tool, WDOIt! (Pinheiro da Silva et al., 2010), provided

by Cyber-ShARE. This is reflected in the top table, through the column named *self description*. The triples for these resources were accessed through the URIs and added to the semantically described collection. Adding semantic descriptions for scientific resources was beyond the scope of this research, in particular because it would have required more time from researchers not allotted at this point. Previous research shows that sharing scientific research on the Semantic Web, in particular data, must consider scientist's concerns for misuse and misunderstanding and, therefore, this should be discussed with scientist's in more detail. Such discussions may enable the use of RDFizers or similar to share research results on the Semantic Web (Gándara and Lapp, 2012).

Having access to posters, presentations and publications is a start to understanding collaboration of this research effort, evident with the high attribution of each. Still, as shown in the bottom table of Table 5.1, not all participants are credited through publications, presentations and posters. Individual researcher and organizational participation was considered at a research effort level, not individual resource level, as attribution was missing from various resources. The list and final count of collaborators, found on the bottom table of Table 5.1 was extracted from attribution in resources as well as from discussions with the research PI. *Found online* identifies those collaborators that are referenced in an online document. If a publication can be accessed through a search on the Web, or perhaps through a researcher's Website, then the organizations and people listed would count under the *Found online* category. *Participation Loss* is the percentage of collaborations that are not found in online documents and do not count for *Found online*. This research effort has 21 individual collaborators. Since only 13 are found in online publications, there is a 38% individual participation loss. Similar was calculated for organizations; only 1 organization of the 5 is identified in online publications resulting in an 80% organizational participation loss.

The ECC Project's Semantic Vocabulary

Due to a lack of semantic descriptions for the original resources, the majority of the properties captured for each resource are collection properties, captured by nodes in the prototype system. This enabled the system to create basic queries and did not require researchers to provide information that was not available or unclear to them at the time, e.g., an ontology to describe resources.

Table 5.2 shows the 21 types used to describe objects within the scientific collection, with 37% defined in the local namespace. Table 5.3 lists the properties used for this research effort. Table 5.3 indicates that 30 properties were added to the local namespace. 25 were added to support nodes within Drupal, i.e., the attributes added to a node that did not have a mapping, and 5 were added during **R**efine to support new relationships between resources.

Table 5.2: ECC Resource Types

SOURCE OF VOCABULARY	NAMESPACE	TYPES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	Person, Organization	2
WGS Geospational Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	SpatialThing	1
RDF Schema	http://www.w3.org/2000/01/rdf-schema	Class	1
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Property	1
OWL	http://www.w3.org/2002/07/owl#	Ontology, ObjectProperty, DatatypeProperty	3
SIOC	http://rdfs.org/soic/ns#	Container, Item	2
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	DataServer, ResourceCollection, ResearchQuestion	3
local server	http://ciserver.leia.com/type/	tool, poster, image, dataset, biblio, program_file, presentation, workflow	8
local server	http://ciserver.leia.com/property#	hasResearchCommitteeMember, hasResearchAdvisor	2
			23

The ECC Project's Scientific Collection

The scientific collection for the ECC project consists of the Web URIs for eddy covariance tools, programs, data and images, posters, presentations and publications that were used or created to conduct research for the eddy covariance cyber-infrastructure. Each resource is characterized as a type in the scientific collection namespace and is added with

Table 5.3: ECC Properties

SOURCE OF VOCABULARY	NAMESPACE	PROPERTIES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	workplaceHomepage, lastName, firstName, homepage	4
WGS Geospatioal Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	long, lat	2
RDF Schema	http://www.w3.org/2000/01/rdf-schema	range, label, isDefinedBy,domain	4
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	type	1
SIOC	http://rdfs.org/soic/ns#	has_space,container_of,about	3
Dublin Core	http://purl.org/dc/terms/	title, subject,rightsHolder,publisher, license,isPartOf,hasPart,description,data, creator,abstract	11
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	sparql_endpoint,research_question_of,research_process, research_logo, longtitle, has_research_question, color, affiliated_with	8
local server	http://ciserver.leia.com/property#	type, revision_timestamp, last_comment_timestamp, language, comment_count, language, format	6
local server	http://ciserver.leia.com/property#	outFigure1, outFigure2, phase, phase_type, isInputTo, hasOutput, inFigure1	7
local server	http://ciserver.leia.com/property#	hasResearchCommitteeMember, hasResearchAdvisor	2
local server	http://ciserver.leia.com/property#	biblio_year, biblio_volume, biblio_type_of_work, biblio_type_name, biblio_type, biblio_tertiary_title, biblio_sort_title, biblio_section, biblio_secondary_title, biblio_publisher, biblio_place_published, biblio_pages, biblio_lang, biblio_doi, biblio_date, biblio_coins, biblio_citekey, biblio_abst_e	18
			66

a **Dublin Core:partOf** relationship to the scientific collection. Resources have collection-based properties identifying the creator (a link to the person's self-describing URI), the organization that owns it (a link to the organization's self-describing URI), location of the resource, e.g., were data was captured, a date a resource was created, e.g., the date a presentation was presented, a description, a title, and abstract. Some of the resources, those that are part of the cyber-infrastructure, have **Dublin Core:subject** properties reflecting the phase of the cyber-infrastructure they pertain to. For example, the PI identified three main phases of the cyber-infrastructure: DESIGN, IMPLEMENT, PROCESS, and sub-phases within those. A resource can have multiple **subject** properties. The datasets that have **Dublin Core:description** properties with hyperlinks to various samples of data have **Dublin Core:seeAlso** relationships to the URL of the hyperlink. Similarly, some resources have properties with hyperlinks to people and to terms in dbPedia. As a result, those resources also have **Dublin Core:seeAlso** relationships with the URL of the hyper-

link. Notice that all of the properties are optional, each resource has triples only for those properties that have values set.

The scientific collection has loaded Web URIs for the people and organizations that collaborated with the research. Comment resources are added as **SIOC: Item** objects that have properties for the creator (a link to the person's self-describing URI), the date, and a relationship with the resource URIs the comment is about.

Properties have been added to the scientific collection's namespace, to document relationships within the cyber-infrastructure, e.g., **outFigure**, **isInputTo**. The properties are used in relationships between resources in the scientific collection to reflect the eddy covariance cyber-infrastructure studied by the ECC project. The SPARQL endpoint for the scientific collection is identified by the **CARP:sparql_endpoint** property and the research related questions are identified through the **CARP:has_research_question** property for the scientific collection.

The ECC scientific collection has 146 URIs that have an associated type and additional properties added during different phases of the methodology. In total, there are 1693 triples created in the semantically described collection for this research effort. Aside from a fixed set of triples that describe the collection and ontology objects (8.67%), 87% of the triples were added during the **C**ollect phase, 1.18% of the triples in the **A**nnotate phase and 3% from the **R**efine phase.

Accessing and Verifying the ECC Project's Scientific Collection

The machine representation for the scientific collection can be accessed in two ways. The self-describing URI for the collection can be accessed to download the scientific collection's ontology. The triples in the self-describing URI can be loaded into a triplestore at a client tool, URIs can be explored and the triples can be analyzed. To verify that correct ontologies are created through the methodology, the ontology RDF, accessible through the self-describing URI was evaluated through an RDF validator (Poveda-Villalón et al., 2012) that scans ontologies for common issues that occur during ontology engineering. There

were minimal issues identified such as suggestions for consistent naming, e.g., avoiding syntactical naming inconsistencies such as ec_tower and ecTower, and consistently adding ontology annotations to elements, e.g., the text 'Eddy Covariance Tower' for the term ec_tower. No ontology-based structure issues were identified such as recursive definitions or using incorrect ontology elements. The implementation of the methodology is capable of only fixing those issues that are created by the methodology, not those imported by external ontologies. In this case study, and in particular because there is little semantic information available, there were no structural issues.

The second method of accessing the scientific collection is through the collection's SPARQL endpoint, accessible through a URL and also set as a property of the scientific collection. The SPARQL endpoint can be queried using SPARQL queries to access some or all of the details of the ECC project. As an example, questions identified by the ECC project PI were transformed into SPARQL queries and executed against the SPARQL endpoint. A graph representation was created, through an HTML page identified in the **P**ublish phase, to show a visual representation of the resources and relationships that result from the queries.

In addition to exposing details specific to the ECC project, applying the research specific SPARQL queries verifies that the scientific collection semantically, i.e., using the appropriate structure, describes the research effort. The queries, specific to the ECC project scientific collection, were executed against the SPARQL endpoints to assure the expected resources and relationships were returned. The ECC triplestore is a relatively small, so assessing query results is still possible through viewing the results of a query. For the ECC project, queries were executed about different phases and sub-phases of the infrastructure returning correct results. Thus, scientific collections, created through CARP, are structurally suited to represent the research of the ECC project case study.

The following questions can be asked about the ECC research effort:
Questions about the cyber-infrastructure:

Question 1 What resources are related to post eddy covariance processing?

Question 2 What resources are related to data validation?

Question 3 What resources are related to data processing and dispersion?

Question 4 What resources are related to quality control and gap filling?

Figure 5.2 shows the result resources and relationships for querying quality control and gap filling. Since resources and relationships are returned, the result graph reflects the queried cyber-infrastructure.

Questions about contributors:

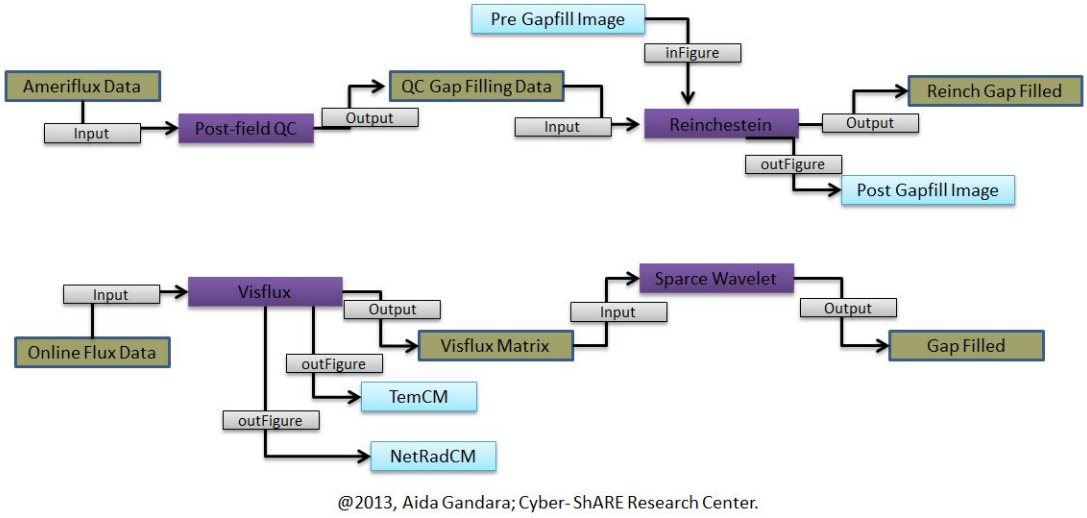

Figure 5.2: Quality Control and Gap Filling resources and relationships documented in the research effort

Question 1 What were collaborative contributions to this research?

Question 2 Who contributed posters?

Question 3 Who contributed programs or tools?

Figure 5.3 shows all the contributors in this research effort and what they did. In this particular view, a lot of information may be a challenge to understand.

Figure 5.4 shows those contributors of posters, a much simpler graph. Both of these questions are SPARQL queries into the same semantically described collection for the case study.

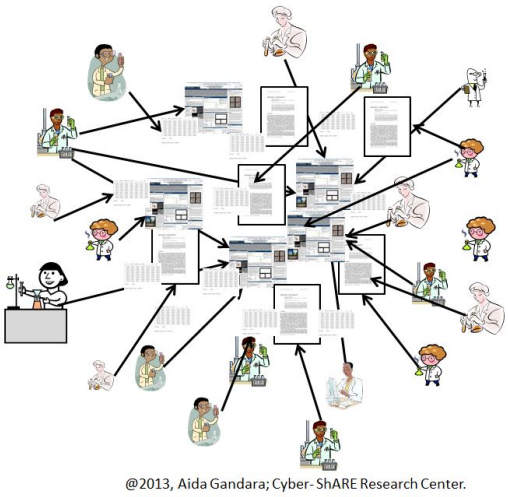

Figure 5.3: A graph representing the results from querying the ECC scientific collection about **all** contributors.

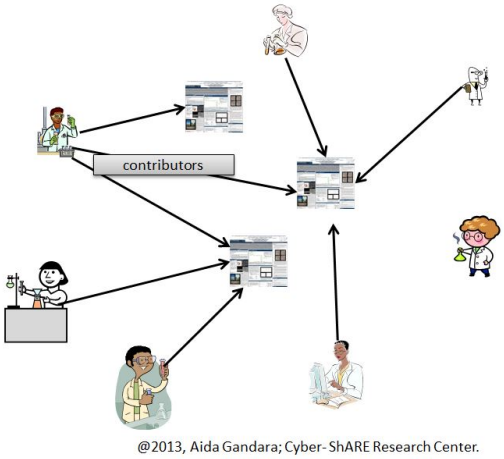

Figure 5.4: A graph representing the results from querying the ECC scientific collection about **only** poster contributors.

5.1.2 Receiver Function Modeling (RFM) Project

Evidence shows that faulting, seismicity and widening are still active geological phenomenon occurring at the Southern Rio Grande Rift(RGR). The RFM project is focused on facilitating the ability to observe certain patterns in interpreting the RGR deformation and extension. The RFM projct computes receiver functions for the RGR area with data gathered from seismic data sources and imaging the crust and mantle structures (Thompson, 2010). Developing the receiver functions and producing 3D images is a detailed process that is only partially described in publications and posters created by the RFM team members, of which there are only a few and with limited accessibility[4].

Overview of the RFM Project

The RFM project researchers have captured over 3.9 Gigabytes of information in about 2251 files, including receiver functions, earth and plot files, scripts and other data. The

[4]The previous paragraph is a summary of the motivation behind the RFM project, collected from resources of the case study.

resources created could be useful in observing details of the Earth as well as programs and datasets that could be valuable for computing earth models. The algorithms for calculating receiver functions and the details for computing and testing the receiver functions are partially documented in posters, presentations and a thesis manuscript, although there is very little documentation concerning the Kriging program or research collaborators. The RFM project is a completed research effort as of December 2011. Although there appear to be no concerns for privacy and versioning, there were challenges in recollection of resource details as well as delays in locating resources because system administrators had relocated files, since the research completed a year ago. In addition, it appeared that the RFM directory had been used for another project since the PI found newer files, not created during research.

The RFM PI was the main contact for discussing and documenting this project. The PI provided a spreadsheet listing data, posters, presentations and programs and provided archives containing a few result receiver function models, scripts, programs and 3D data. RFM scientists acquired data from two different data sources to conduct the analysis: 1) data was acquired from seismic data providers through a download; 2) receiver functions and Kriging scripts were acquired from another researcher. The acquired scripts were customized by the PI to generate receiver functions for 144 receiver stations in the Southern RGR and Kriging scripts were customized to convert result 1 Dimensional (1D) data to 3 Dimensional (3D) data. Any files owned by data providers were not included in the documented resources. Other researchers in Cyber-ShARE have reused the result receiver function data as well as Kriging scripts, aside from this, there is no additional request to share this research with another community. Resources were evaluated for semantic structure and attribution. The questions for this research were initially focused on identifying good receiver function models versus bad ones. Unfortunately, only a good model was provided in the data files. Instead, research questions focus on describing the receiver function and Kriging process as well questions concerning the receiver stations that relate to receiver function data.

RFM Project Resources

After an initial evaluation of the files and spreadsheet provided by the PI, 600 resources were considered for documenting this research effort. Table 5.4 shows two tables, the top is a summary of where the 600 resources from this case study were located at the beginning of the documentation; the bottom table shows collaborator information captured from the different resources and discussions with researchers, showing collaborator qualities at the beginning of the documentation. Research information for this research effort was found in directories on a researcher's laptop but mostly on a secured server accessible to the research group. The fact that this research ended over a year ago introduced challenges to collecting research resources, such as difficulty in recalling details on how to reproduce results, problems locating files related to the research because they had since been moved, and finding research files from subsequent research interspersed in the directories.

As shown in the top table of Table 5.4, the majority of the files for this research are images and data. Some of the files were related collections from applying a program and generating a fileset of hundreds of files. Archives were created from the filesets to facilitate downloads. To make the various resources, including the fileset archives, accessible on the Web they were uploaded to a server, assigned a URI and added to the semantically described collection. Uploading the resources also assured that the files could not be misplaced or merged with other projects. *Online* resources include Webpages for vendor tools used in the project and for dataset providers to obtain the datasets for the SGR region. In the cases where data was not available but was useful in describing research, *Neither* was used to account for those resources. For example, the Kriging scripts process is used by other research efforts, thus documenting it is useful, however, there is no sample data in the RFM data fileset. The row labeled % shows a summary of resource locations for this research with 1.33% available online and almost 98% on local server drives and less 1% of the resources of interest not found.

Table 5.4: Two tables summarizing the state of resources at the beginning of documenting the RFM project case study. The top table shows the different resources, where they were located, their attribution and self-description. The bottom table shows the collaborator information calculated from the resources, including self-description, the collaborations found in some online document and participation loss.

RESOURCE TYPES	Local/ Personal Drive	Online	neither	TOTAL	attribution	%attribution	self description
images / animations	144	0	0	144	0	0.00%	0
workflow	0	2	0	2	0	0.00%	2
programs	4	0	0	4	0	0.00%	0
tools	0	3	0	3	3	100.00%	0
presentation	2	0	0	2	2	100.00%	0
posters	5	0	0	5	5	100.00%	0
publications	0	1	0	1	1	100.00%	0
data	433	2	4	439	0	0.00%	0
TOTAL	588	8	4	600	11	1.83%	2
%	98.00%	1.33%	0.67%				
stations	143	0	0	743			
%w stations	98.38%	1.08%	0.54%				

COLLABORATORS	TOTAL	self description	found online	participation loss
people	8	0	4	50.00%
organizations	5	0	1	80.00%
TOTAL	13	0	5	

Each resource was evaluated for a reference to the owner or creator. This was a challenge, as seen in the 1.85% *attribution* average shown in Table 5.4. Posters, publications and presentations were created for sharing and have a high attribution, as seen by the attribution columns for each. Images, workflows, programs and data do not have attribution. The programs used for this research were provided by another scientist with permission to edit and reuse. The exchange was based on a verbal agreement. There is no attribution in these files to either the original creator or members of this research team. The RFM project utilized online tools and those were all considered as attributed since there is Web-accessible access to a vendor Website. Overall, less that 1.5% attribution was found in the files for this research.

An evaluation was performed on the resources to identify which were self-describing or had microformats. Similar to the ECC project, only two resources had semantic representations because they are OWL workflows.

Providing access to posters, presentations and publications is a start to understanding collaboration of this research effort, this is evident with the 100% attribution of each of these. Still, as shown in the bottom table of Table 5.4, not all participants are credited through those three resource types. The list and final count of collaborators, found on the bottom table of Table 5.4 was extracted from attribution in resources as well as from discussions with the research PI. Similar to the ECC project, *Participation Loss* was measured based on whether a collaborator was referenced in an online document. This research effort has 8 individual collaborators. Since only 4 are found in an online publication, there is a 50% individual participation loss. Similar was calculated for organizations; only 1 organization of the 5 is identified in the online publication resulting in an 80% organizational participation loss.

The RFM Project's Semantic Vocabulary

The majority of the properties captured for the resources of this scientific collection were captured from the collection properties defined in the nodes. One benefit of the method-

ology was the ability to use defaults until more semantic information is available. This feature enabled users to create basic queries and did not require researchers to provide information that was not available or unclear to them at the time, such as an ontology.

In total, there were 27 properties added to the local namespace. 22 were added to support nodes within Drupal and 5 were added during Refine, to support new relationships and to load more details about URIs, e.g., the stations. There are 22 types used to describe objects within the research description, as listed in Table 5.5, with (41%) defined in the local namespace. Table 5.6 lists the properties used for this research effort. There are 63 properties utilized in the description of this research effort with 48% defined in the local namespace. Since semantically described collections are exposed as SPARQL endpoint, software agents can query the endpoint and explore the vocabulary, listed in Tables 5.5 and 5.6.

Table 5.5: RFM Resource Types

SOURCE OF VOCABULARY	NAMESPACE	TYPES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	Person, Organization	2
WGS Geospational Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	SpatialThing	1
RDF Schema	http://www.w3.org/2000/01/rdf-schema	Class	1
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Property	1
OWL	http://www.w3.org/2002/07/owl#	Ontology, ObjectProperty, DatatypeProperty	3
SIOC	http://rdfs.org/soic/ns#	Container, Item	2
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	DataServer, ResourceCollection, ResearchQuestion	3
local server	http://ciserver.leia.com/type/	tool, poster, image, dataset, biblio, program_file, presentation, workflow, station	9
			22

The RFM Project's Scientific Collection

The scientific collection for the RFM project consists of the Web URIs for the different resources provided by the PI. Some of the resources are archives and other files, but each is accessible from the description within the scientific collection. Resources are characterized as types defined in the scientific collection namespace and resources are added to the scientific collection with the **Dublin Core:partOf** relationship. Resources have properties

Table 5.6: RFM Properties

SOURCE OF VOCABULARY	NAMESPACE	PROPERTIES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	workplaceHomepage, lastName, firstName, homepage	4
WGS Geospatioal Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	long, lat	2
RDF Schema	http://www.w3.org/2000/01/rdf-schema	range, label, isDefinedBy,domain	4
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	type	1
SIOC	http://rdfs.org/soic/ns#	has_space,container_of,about	3
Dublin Core	http://purl.org/dc/terms/	title, subject,rightsHolder,publisher, license,isPartOf,hasPart,description,data, creator,abstract	11
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	sparql_endpoint,research_question_of,research_process, research_logo, longtitle, has_research_question, color, affiliated_with	8
local server	http://ciserver.leia.com/property#	type, revision_timestamp, last_comment_timestamp, language, format, original creator	6
local server	http://ciserver.leia.com/property#	phase, phase_type, isInputTo, hasOutput	4
local server	http://ciserver.leia.com/property#	network, station_id, elevation	3
local server	http://ciserver.leia.com/property#	biblio_year, biblio_volume, biblio_type_of_work, biblio_type_name, biblio_type, biblio_tertiary_title, biblio_sort_title, biblio_section, biblio_secondary_title, biblio_publisher, biblio_place_published, biblio_pages, biblio_lang, biblio_doi, biblio_date, biblio_coins, biblio_citekey, biblio_abst_e	17
			63

defined for all resources of the collection, e.g., creator, owner, location, data, title, abstract, etc. The resources that have properties with text containing hyperlinks, have explicit **Dublin Core: seeAlso** relationships to the URL of each hyperlink. For example, the resource describing RGR data downloaded from a data provider has a **Dublin Core: description** property with instructions for accessing and downloading data from data providers, the description has a hyperlink to the data provider site. The prototype system generates a **Dublin Core:seeAlso** property from the resource to the provider site. The scientific collection has comments as **SIOC:Item** objects related to resources through the **SIOC:about** relationship.

A self-description was created for the stationlist resource where the stations have individual hashed URIs. The self-describing URI for the stationlist is loaded into the scientific collection. The receiver function resources are identified as part of the receiver function process with a **Dublin Core:subject** property. Similar is done for the resources in the Kriging process. Finally, to describe the relationships between the resources and data of the Kriging and receiver function processes, properties are added to the scientific collection namespace and relationships between data and programs are added to the collection. The SPARQL endpoint for the scientific collection is identified by

the **CARP:sparql_endpoint** property and the research related questions are identified through the **CARP:has_research_question** property for the collection.

The scientific collection contains 219 URIs that have an associated type and additional properties. Notice that this number might appear lower than expected because much of the data was added to archives. The result scientific collection has 1874 triples with 38% added during **C**ollect phase of the case study; 2.51% of the total triples were added during the **A**nnotate phase; and 55% of the triples added to the scientific collection during the **R**efine phase.

Accessing and Verifying the RFM Project's Scientific Collection

The machine representation for the scientific collection can be accessed in two ways, through a self-describing URI and through a SPARQL endpoint, both created automatically by the prototype system. The triples in the self-describing URI can be loaded into a triplestore at a client tool where URIs can be explored and triples analyzed. To verify that the RFM project's ontology is an acceptable ontology, the RDF was evaluated through an RDF validator (Poveda-Villalón et al., 2012) that scans ontologies for common issues that occur during ontology engineering. No ontology-based structure issues were identified such as recursive definitions or using incorrect ontology elements. The SPARQL endpoint is accessible through a URL that accepts sparql queries to access some or all of the details of the RFM project. The questions identified for the RFM project were transformed into SPARQL queries and executed against the SPARQL endpoint. An HTML view of the results are returned as a graph, to show a visual representation of the resources and relationships that result from the queries.

In addition to exposing details specific to the RFM project, applying the research specific SPARQL queries verifies that the scientific collection semantically, i.e., using the appropriate structure, describes the research effort. The queries, specific to the RFM project ontology, were executed against the SPARQL endpoints to assure the expected resources and relationships are returned. The RFM triplestore is a relatively small ontology, so

assessing query results is still possible through viewing the results of a query. For the RFM project, queries were executed about different phases and sub-phases of the infrastructure returning correct results. Thus, scientific collections, created through CARP, are structurally suited to represent the research of the RFM project case study.

The following questions can be asked about the RFM research effort:

Question 1: What does the seismic structure look like at lat 32.01 and long -106.43?

Question 2: Where are the receiver function stations located?

Question 3: What are the resources used in the Kriging Moho analysis?

Question 4: What is known about the Kriging Moho program?

Question 5: What is the receiver function process?

Figure 5.5 and Figure 5.6 show two examples of the results obtained from asking Question 2. Figure 5.5 shows an RFM image with receiver station locations for the Southern Rio Grande Rift whereas Figure 5.6 is a dynamically generated map from querying the triplestore for the lat and long of the receiver function **stations**.

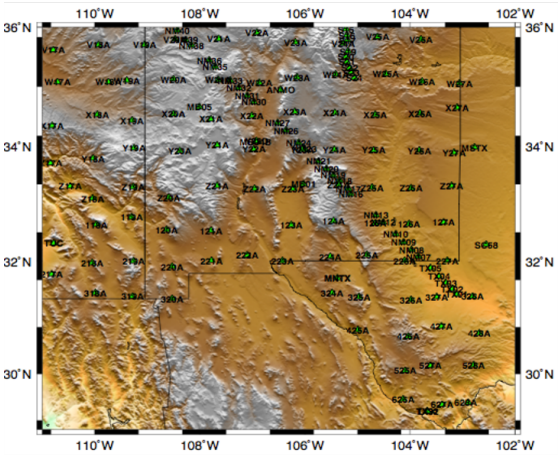

Figure 5.5: Image showing all receiver function locations

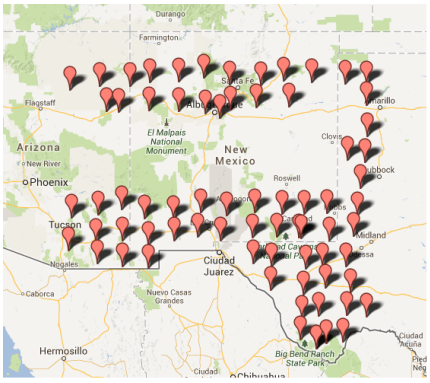

Figure 5.6: Map generated from self-describing a station table

5.1.3 Constraint Optimization (CO) Project

A key problem in geophysics is the use of multiple information sources in order to more accurately determine physical properties of the Earth. One example is the of teleseismic P-wave receiver functions and surface wave dispersion velocities to estimate a one dimensional (1D) Earth structure. The CO project has developed an optimization strategy that incorporates physical bounds as an explicit constraint over model parameters. The optimization occurs over the 1D problem and subsequent code, produced by the RFM project, is used to produce the 3D model. This research has worked over several synthetic models as a basis for validating the optimization called the primal-dual interior-point (PDIP) method (Sosa, 2012). The details of the process and comparison to the synthetic models has been described in several posters and the dissertation from the principal investigator. However, the details of the programs used and generated by this research as well as its reuse of a previous research effort's results are not generally available. The posters and presentations for this research are available on a Website but posters and presentations lack explanations of and access to the programs that could facilitate reproducing results[5].

Overview of CO Project

The CO project has over 200 Megabytes in 500 files that were used or created to represent different synthetic and Earth models for this research effort. The studies and the PDIP code itself could prove useful to other scientists interested in validating or reusing the technique. The details of the algorithms and techniques are published in posters and presentations, although there are details about the research that are not included. The CO project recently ended in December of 2012. There were no mentioned concerns for privacy or issues with locating files.

The main contact for this research was the CO project PI. The initial spreadsheet had a listing of posters and presentations as well as the code and data created or used during the

[5]The previous paragraph is a summary of the motivation behind the CO project, collected from resources of the case study.

research. The CO scientists acquired synthetic data sets for receiver function and surface wave data and these were used to initially validate the optimization. In addition, programs were downloaded and used to conduct the joint inversion, optimizations and 3D interpolation. The filesets for the two tools used in the study were excluded from the resources considered for the documentation of this research, however, the filesets were referenced in the description of datasets where appropriate. A list of collaborators was collected from the PI and from scanning resources. There were no discussions about semantic vocabularies, aside from an initial introduction discussing the goals of scientific collections. There are ongoing efforts to submit publications on this research, however, there are no immediate plans to share additional details, e.g., the PDIP code, of this work with other organizations. Resources were evaluated for semantic structure. In addition, the research PI felt that the different models were redundant and it was not necessary to document all the models to describe this research. It was decided to document one if the synthetic models (the Archean model) and an SGR model resulting from a collaboration with the RFM project. The questions identified for this research related to the Archean model, the PDIP outputs and the relationships between resources in both the RFM and CO projects.

CO Project Resources

After an initial evaluation of relevant files provided by the PI, 174 resources were considered for documenting the details of case study. Table 5.7 displays two tables, the top shows a summary of where the resources from this case study were located at the beginning of the documentation process; the bottom table shows collaborator information captured from the different resources and discussions with the PI, showing collaborator qualities at the beginning of the documentation process. Research resources for this research effort were found in directories on a secured server and accessible through the research PI's Webpage. Almost 78% of the research resources were found in directories on a secure server belonging to the research group, including images, programs and data. These were uploaded to a Cyber-ShARE server and assigned URIs. To avoid losing information, as occurred with the

RFM project, the additional synthetic models were also uploaded to a Cyber-ShARE server but were not added to the collection for the case study. The 15.15% found *Online* included tools, data, workflows, presentations, posters and publications that are accessible on the Web. Two codesets were borrowed from other sources, one to compute joint inversion and the other to compute seismic travel times. Each tool was considered a single codeset since they are downloaded as a single resource. Also the research PI has an online Webpage where posters and presentations are available. Finally, the research PI downloaded synthetic data used to validate the studies but the source of this data was not documented. For the intentions of this research documentation the source Archean data was *Neither* available online or on a local server.

Table 5.7: Tables showing resources and collaborators of the CO project

RESOURCE TYPES	Local Server	Online	neither	TOTAL	attribution	attribution %	self description
image / animation	16	0	0	16	0	0.00%	0
workflow	0	2	0	2	0	0.00%	2
program	44	0	0	44	40	90.91%	0
tool	0	2	0	2	2	100.00%	0
presentation	0	9	0	9	9	100.00%	0
poster	0	6	0	6	6	100.00%	0
publication	0	1	0	1	1	100.00%	0
data	84	7	3	94	0	0.00%	0
TOTAL	144	27	3	174	58	33.33%	2
%	82.76%	15.52%	1.72%				

COLLABORATORS	TOTAL	self description	found online	participation loss
people	10	0	9	10.00%
organizations	7	0	6	14.29%
TOTAL	17	0	10	

Each resource was evaluated for a reference to the owner or creator. This was a challenge, as seen in the 1.85% *attribution* average shown in Table 5.7. Posters, publications and presentations were created for sharing, for the most part, and have a high attribution as seen by the attribution columns for each. Although images and workflows might as well, researchers did not capture attribution for these. No attribution was captured for programs and data either. The code provided by the researcher, computing the new PDIP algorithm, was attributed through only one file, a README, included in the directory. As a result, it was given a full attribution value, for the 40 files included, because the entire codeset is being shared as a single archive. The joint inversion tool exposed a different scenario. After downloading the codeset, the PI modified some of the files. The files are identified as local and without attribution due to the changes. The program itself was still attributed to the original creator. This research had an overall attribution of about 33%.

Providing access to posters, presentations and publications is a start to understanding collaboration of this research effort, this is evident with the 100% attribution of each of these. The bottom table in Table 5.7 shows that not all participants are credited through those three resource types. The list and final count of collaborators was extracted from attribution in resources as well as from discussions with the research PI. Attribution in programs, was handled differently. The owners name was only added if the researcher was mentioned in a poster, presentation or publication. Similar to the ECC project, *Participation Loss* was measured based on whether a collaborator is referenced in an online document. This research effort has 10 individual collaborators. This research, to date, only has one publication, however, the posters and presentations are online so names listed in those documents counted in the *found online* column. For this research, there is only a 10% individual participation loss. Similar was calculated for organizations; 6 of the 7 listed organizations are identified in online documents resulting in a 14% organizational participation loss.

The CO Project's Semantic Vocabulary

The majority of the properties captured for the resources of this scientific collection were captured from the collection properties defined in the nodes of the prototype system.

In total, there are 21 types used to describe objects within the scientific collection, as listed in Table 5.8, with 38% defined in the local namespace. Table 5.9 lists the properties used for this research effort. There are 64 properties utilized in the scientific collection with 48% defined in the local namespace. Since the scientific collection is exposed as a SPARQL endpoint, software agents can query the endpoint to explore the vocabulary.

Table 5.8: CO Resource Types

SOURCE OF VOCABULARY	NAMESPACE	TYPES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	Person, Organization	2
WGS Geospational Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	SpatialThing	1
RDF Schema	http://www.w3.org/2000/01/rdf-schema	Class	1
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	Property	1
OWL	http://www.w3.org/2002/07/owl#	Ontology, ObjectProperty, DatatypeProperty	3
SIOC	http://rdfs.org/soic/ns#	Container, Item	2
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	DataServer, ResourceCollection, ResearchQuestion	3
local server	http://ciserver.leia.com/type/	tool, poster, image, dataset, program_file, presentation, biblio, workflow	8
			21

The CO Project's Scientific Collection

The scientific collection for the CO project consists of Web URIs for the different resources provided by the PI. Some of the resources are archives of data models. The resources are accessible through the URI. Resources are characterized as types defined in the local CO collection namespace and resources are added with the **Dublin Core:partOf** relationship

Table 5.9: CO Properties

SOURCE OF VOCABULARY	NAMESPACE	PROPERTIES	TOTAL
FOAF ontology	http://xmlns.com/foaf/0.1/	workplaceHomepage, lastName, firstName, homepage	4
WGS Geospatioal Ontology	http://www.w3.org/2003/01/geo/wgs84_pos#	long, lat	2
RDF Schema	http://www.w3.org/2000/01/rdf-schema	range, label, isDefinedBy,domain	4
RDF	http://www.w3.org/1999/02/22-rdf-syntax-ns#	type	1
SIOC	http://rdfs.org/soic/ns#	has_space,container_of,about	3
Dublin Core	http://purl.org/dc/terms/	title, subject,rightsHolder,publisher, license,isPartOf,hasPart,description,data, creator,abstract	11
CARP	http://cybershare.utep.edu/ontology/carp-ns.rdf#	sparql_endpoint,research_question_of,research_process, research_logo, longtitle, has_research_question, color, affiliated_with	8
local server	http://ciserver.leia.com/property#	biblio_year, biblio_volume, biblio_type_of_work, biblio_type_name, biblio_type, biblio_tertiary_title, biblio_sort_title, biblio_section, biblio_secondary_title, biblio_publisher, biblio_place_published, biblio_pages, biblio_lang, biblio_doi, biblio_date, biblio_coins, biblio_citekey, biblio_abst_e	17
local server	http://ciserver.leia.com/property#	type, revision_timestamp, last_comment_timestamp, language, comment_count, language, format	7
local server	http://ciserver.leia.com/property#	outFigure1, phase, phase_type, isInputTo, hasOutput	5
local server	http://ciserver.leia.com/property#	hasResearchCommitteeMember, hasResearchAdvisor	2
			64

to the collection. Resources have some properties that are consistent for all resources in a collection, e.g., creator, owner, location, data, title, abstract, etc., however, all properties are optionally supplied and can not be assumed within the collection. Resource properties would have been additional to any other semantic properties harvested from the resources, but most of the resources did not have self-describing URIs or embedded RDF that could be harvested. Triples are added for resources that have properties containing text with hyperlinks, i.e., there are explicit relationships for those resources to the URLs in the hyperlinks. Adding the properties from the hyperlinks to the scientific collection is part of the **A**nnotate phase of the methodology. The collection has comments as **SIOC:Item** objects that contain observations captured from discussions and the resources provided by the PI about some of the result resources, e.g., to describe the PDIP output table. Comments are related to resources through the **SIOC:about** relationship. The programs and data for the models, e.g., the Archean synthetic model and the RGR model, are

identified with the **Dublin Core:subject** property. Relationships exist between data and programs to represent the process flow. Properties to support these relationships are defined in the namespace for the collection ontology. Since this research effort relies on processes and data from the RFM project, the self-describing URIs for the Kriging and SRG datasets created during the RFM project were loaded into this scientific collection. The SPARQL endpoint for the scientific collection is identified by the **CARP:sparql_endpoint** property and the research related questions are identified through **CARP:has_research_question** properties for the collection.

The CO project scientific collection identifies 91 URIs that have an associated type and additional properties. Notice that this number might appear lower than expected because much of the data was added to archives. The result scientific collection has 1064 triples with approximately 75% added during **C**ollect phase of the case study; approximately 5% of the total triples were added during the **A**nnotate phase; and approximately 7.5% of the triples added to the collection during the **R**efine phase.

Accessing and Verifying the CO Project's Scientific Collection

The machine representation for the scientific collection can be accessed in two ways, through a self-describing URI and through a SPARQL endpoint, both created automatically by the prototype system. The triples in the self-describing URI can be loaded into a triplestore at a client tool and URIs can be explored and the triples can be analyzed. To verify that the CO project's ontology is an acceptable ontology, the RDF was evaluated through an RDF validator (Poveda-Villalón et al., 2012) that scans ontologies for common issues that occur during ontology engineering. No ontology-based structure issues, e.g., recursive definition or incorrect ontology elements, were identified.

The SPARQL endpoint is accessible through a URL that accepts sparql queries to access some or all of the details of the CO project. The questions identified by the CO project PI were transformed into SPARQL queries and executed against the SPARQL endpoint. The HTML views created during the ECC project were leveraged to show a visual representation

of the resources and relationships that result from the queries.

In addition to exposing details specific to the CO project, applying the research specific SPARQL queries verifies that the scientific collection semantically, i.e., using the appropriate structure, describes the research effort. For the CO project, queries were executed about the Archean model, PDIP outputs and creators of resources. Questions about creators are meant to expose the collaboration that occurred between two scientific collections (RFM and CO) showing how two SPARQL endpoints can be queried to find similarities. The queries, specific to the CO project ontology, were executed against the SPARQL endpoints to assure the expected resources and relationships are returned. The CO triplestore is a relatively small ontology, so assessing query results is still possible through viewing the results of a query. Thus, scientific collection, created through CARP, are structurally suited to represent the research of the CO project case study.

The following questions can be asked about the CO research effort:

Question 1: What are the datasets for the Archean model?

Question 2: What information is there about the PDIP outputs of the Archean model?

Question 3: What information is there about the PDIP input data for the Rift model?

Question 4: Who created the kriging 3D interpolation program?

Figure 5.7 shows the information returned for PDIP outputs of the Archean model. Essentially, the returned information is the PDIP output table and the comments. Subsequent queries about the comments can return more information like the author and the date.

5.1.4 Summary

The main goal of the case studies is to build scientific collections of research related results. Through the case studies the research was able to capture similarities and differences in the challenges and benefits of documenting scientific research.

All three research efforts had the majority of resources on local or personal drives (92%)

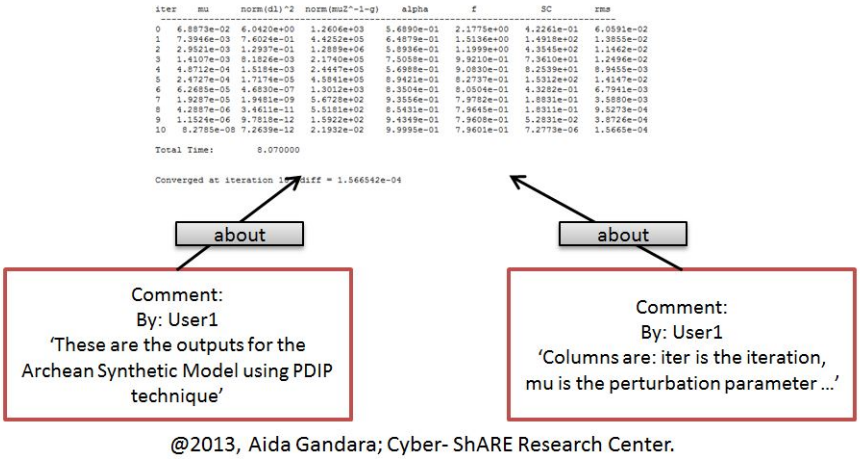

Figure 5.7: PDIP output for the Archean model, including comments.

and two of the research efforts had attribution in less than 2% of the resources produced by the research effort. The third research effort had attribution in less than half (33%) of the resources produced. CARP causes researchers to consider Web-accessibility and attribution, two topics that are not currently part of a systematic research process. To support the methodology, Cyber-ShARE now has a published attribution and licensing policy[6]. Consequently, some resource attribution was added to case study resources and some resources were uploaded to the prototype system but the time to retroactively perform this for thousands of resources was not feasible for the case studies. To save time, CARP focused on documenting resources for selected research related questions.

Few resources and collaborators have semantic descriptions (less than 1% for all three research efforts) thus, as resources were initially added to a scientific collection, having a predefined vocabulary aided in describing a default scientific collection. This was not sufficient to describe more detailed qualities, i.e., the station list that was described for the RFM project, but it was enough to show researchers a list of resources related from their research and to query those resources based on default properties. The research found that semantic vocabularies were never a topic initiated by the case study PIs for describing

[6]Cyber-ShARE's attribution and licensing policy can be found at: http://cybershare.utep.edu/content/cyber-share-acknowledgment

research results. Although the case study researchers were not ready to integrate with other organizations, there is interest, e.g., the ECC project plans on integrating curated eddy covariance datasets in a database created by the group and with the Fluxnet network.

The phases of the methodology guided the systematic documentation of resources to create scientific collections, however, there were challenges. For example, the RFM project encountered issues with locating files that delayed the case study and the ECC project encountered issues with privacy that limited the files that were uploaded to the prototype system. The RFM project also encountered an issue in answering a research related question because needed data was not preserved. Nevertheless, with the methodology, scientific collections were generated for each research effort, as were machine understandable queries. Queries were identified to ask about data processing, about collaborators and other research details. The research even exhibited federated queries to demonstrate the ability to query multiple scientific collections to find similar resources used across two different research efforts.

5.2 A Survey on Scientific Collections

The process of documenting three case studies provided insight into documenting scientific research, from research results on personal drives and servers to Web-accessible resources within the context of a scientific collection. Scientific collections demonstrate documenting scientific research on the Semantic Web for machine understanding, enabling machines to process information uniformly. Future work for this research will consider approaches for processing and making scientific collections more understandable to people, and, in particular, within specific scientific domains. Currently, the prototype for CARP can produce some basic graphs, lists of research related resources and a query form to exhibit the benefits of using scientific collections to document scientific research. To assess whether scientists see a benefit to this approach of sharing scientific results, a survey was conducted with 27 researchers from Cyber-ShARE. Most of the participants surveyed are working under

similar situations as the case study team members; they have conducted or are conducting scientific research, they produce various research resources including data, posters, publications, images, etc., some might have standard approaches for sharing their data but most do not, and they will need to publish their work for others to potentially reuse. In addition, they work for the Cyber-ShARE Center that needs automated access to the collection of their work to include the Center's research outcomes.

To conduct the survey, the prototype environment was presented and each survey participant was asked to respond to a set of questions. To exhibit what was captured for the case studies, a Webpage was shown with all the information collected for a research effort, including a list of all the resources, the people that collaborated with the research effort, and information about how to provide attribution for the research effort. In addition views were shown with graphs resulting from queries, e.g., to show relationships between collaborators and resources, relationships between resources and relationships between tools and data. The questions asked whether survey participants believed that this method of documenting scientific research could : 1) increase accessibility to related research information that was produced during scientific research; 2) increase understanding of research resources by accessing views into the information; 3) convey the contributors and attribution of a research effort; and 4) increase access to notes and discussions that could supplement understanding research.

Figures 5.8- 5.11 show the results of the survey. Questions were asked on a scale of agreement, including: strongly agree, agree, disagree, strongly disagree. For each topic area, an initial question gauged the participants opinion about the topic in general. Figure 5.8- 5.10 show 100% agreement that accessibility, understanding and attribution of resources are important, respectively, with variation between agree and strongly agree and 0% disagreement. For accessibility (Figure 5.8), the majority of participants strongly agree or agree (30%, 60%) that this methodology will increase accessibility of research resources. The expectation that this methodology will increase understanding of individual resources, shown in Figure 5.9, is mainly agreed upon (53%) by participants with strongly agree

(21%) and disagree (18%) almost even. Only 4% strongly disagree. The expectation is more favorable that the methodology will increase understanding of the overall research; 63% participants agree, with strongly agree (15%) and disagree (11%) almost even and 4% strongly disagreeing. When participants were asked if they felt they could identify the contributors of the research effort, the majority strongly agree (33%) and agree (59%), with only 7% disagreeing and no strong disagreement. In addition, when asked if they understand how to attribute the research effort, the majority (59%) agree, however strongly agree (22%) and disagree (18%) are near even. Having access to notes and discussion, shown in Figure 5.11, is primarily strongly agreed (44%) and agreed (41%) upon, however, as opposed to the other three topics where there is 0 disagreement, the importance of notes and discussions has a 15% disagreement. Nevertheless, 70% of participants agree that this methodology could increase access to notes and discussions, with a 19% strong agreement and 11% disagreement.

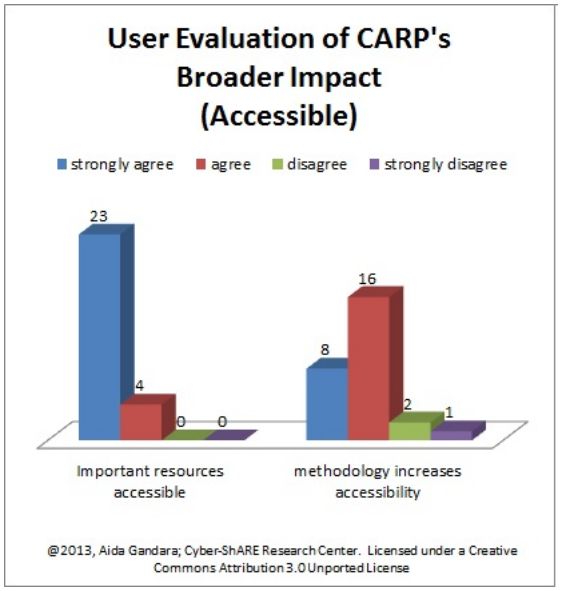

Figure 5.8: Responses about accessibility of resources

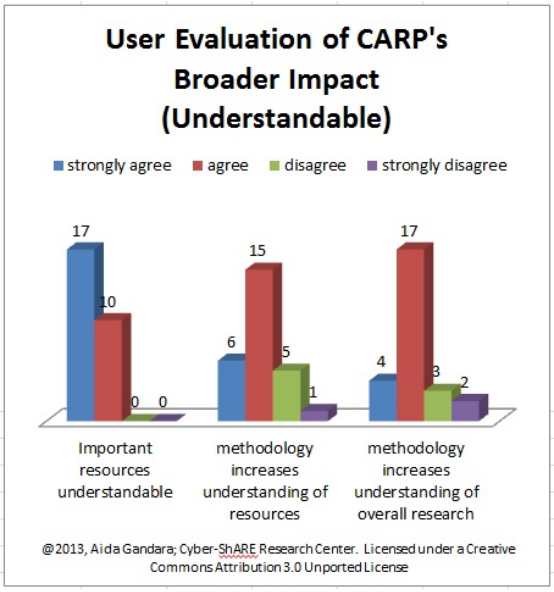

Figure 5.9: Responses about understanding of resources

Due to the increased availability in social tools, e.g., facebook and twitter, and with findings suggesting that notes and discussions may help with understanding scientific re-

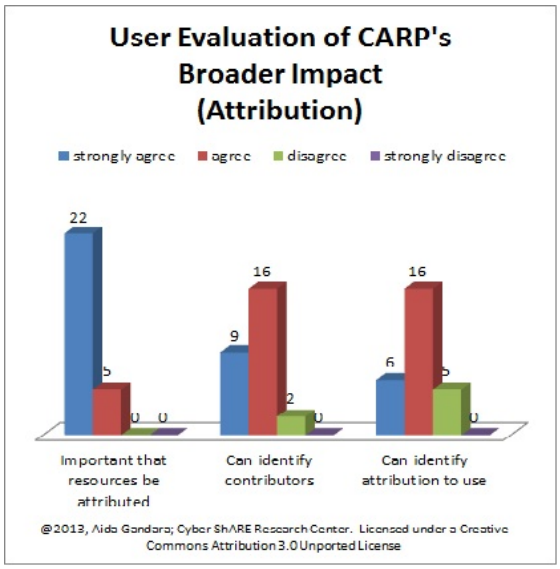

Figure 5.10: Responses about attribution of resources

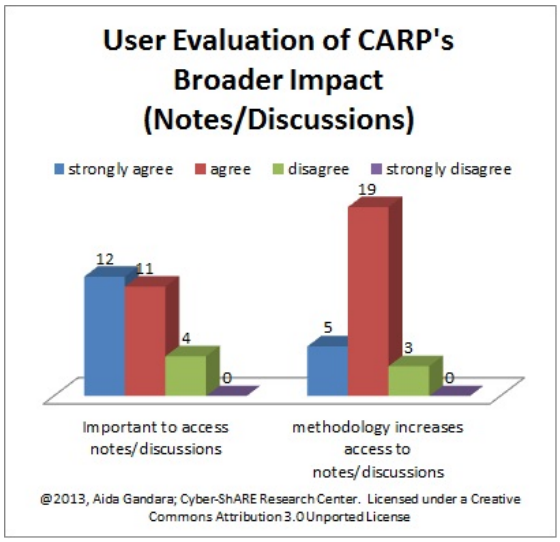

Figure 5.11: Responses about notes and discussions about resources

search (Gándara et al., 2011a), additional questions were included to assess researcher's opinions about adding more notes and discussions to scientific collections. Although in general, the results were in agreement to include them, there were more objections than for the other three topics (attribution, accessibility, understanding), as seen in Figure 5.11. Support for a more elaborate collaborative environment to embed more comments and discussions in scientific collections requires further investigation into what collaboration details to capture and when or how to share them.

Chapter 6

Discussion

The research introduced challenges and lessons learned when documenting scientific research. The chapter discusses the more noteworthy observations.

6.1 Research stages

The research found that the stage of a research effort affects the documentation process. The process to document scientific research is not predefined and can be more complicated or delayed for one research effort versus another. The ECC project showed that if a project is in the crux of research, there is more indecision and concern for what can be shared, requiring that the documentation process occur at a slower and more cautious pace. Publishing some resources requires ongoing considerations as to meaning of a resource and privacy. Such concerns are understandable since findings and conclusions are still being discussed.

Delaying documentation of research results created multiple challenges for the RFM project. Since a year had passed since the conclusion of the research effort and the PI was already working on other projects, there were delays in recollecting details about the results. Files were misplaced and new files were located in the fileset. These details delayed the documentation process and even limited the amount of time that the researcher could spend describing relationships and adding annotations.

There were benefits in the recent conclusion of the CO project. Since this research completed recently, the PI is currently focused on publishing details about the work for a journal. As a result, three benefits were observed: 1) the research PI had a clearer sense

about what was and was not of interest for describing the project; 2) the research PI was certain of the whereabouts of all files and data reducing the amount of time and interactions needed to follow CARP; and 3) the research PI was certain as to privacy concerns for research related resources and only shared resources that were ready for publication.

6.2 Lack of evidence

Documenting scientific research could benefit from planning data capture to reflect questions that should be answered. One finding identified in the case studies is that lack of evidence limits what can be documented and, therefore, questions that can be answered about a research effort. For the RFM project, the lack of evidence prevented answering a question considering 'good' versus 'bad' models based on receiver functions because there was only one 'good' model in the provided fileset. The research suggests that result data be preserved with research questions in mind, i.e., by capturing data that justifies the results. By considering the evidence that will be shared while conducting research, not just successful models, the RFM project may have maintained unsuccessful models ('bad' models) as well as the 'good' model maintained for publication.

6.3 Attribution and licensing policy

Attribution and licensing should be either an automated step or required manual policy. In all case studies, most of the resources were missing attribution and researchers did not have the time to go back and rectify this. Moreover, for licensing, researchers were unsure as to an appropriate licensing policy and had to seek the answer from their supervisors or the Center. As a result, attribution and licensing was captured as a property in the scientific collection but was not included directly in the resources. Attribution and licensing that is not embedded in the resource is lost if the file is downloaded. The question at this point is whether the capture of licensing and attribution should have occurred earlier or

more time needs to be allotted during scientific documentation to add it. The research recommends that adding attribution and licensing be incorporated earlier in the research process. Each case study produced a high number of files, thus, there are too many files to modify. A better measure would be to capture this as resources are created, either through automation where tools prompt for this information or manually.

6.4 RDFization or Self-description

One issue to consider with a methodology such as CARP and working over the Semantic Web is how the meaning for Web resources is obtained. The role of providing self-describing URIs or even embedded microformats is a function of the server hosting Web resources, but not all servers provide such functionality. RDFizers are useful in cases where structured information, in RDF, is needed from a resource that does not have a semantic description. There are issues with RDFizers in terms of authenticity and consistency. If a client tool RDFizes a Web resource, there is no guarantee that this is the accepted or correct meaning, unless the RDFizer is provided by the provider of the Web resource. Moreover, if two different clients use different RDFizers for the same Web resource, there is no guarantee that the two clients obtain the same meaning. CARP promotes self-describing URIs over RDFization when possible to assure that meaningful descriptions are consistent and authentic. The CO project exhibited this benefit of self-describing URIs by reusing some from the RFM project. Since self-describing URIs are individually accessible, loading their meaning in the CO project is identical to that of the RFM project, regardless of the collection a URI is added to. The effort to make self-describing URIs for resources accessible for a research effort is only required once, i.e., on the server that dereferences the URI.

6.5 Publication tools and self-descriptions

To facilitate documenting scientific collections, tools can be used to capture information that is ready to be shared. Resources that are generated by data capturing or data publishing tools can facilitate the documentation process by providing templates to structure content that automatically builds a self description. For example, if users are creating posters, a template can be used where different parts of a poster have structured vocabulary embedded. In addition, resources that are known on the Semantic Web can be linked, e.g., links to authors, published datasets or videos. Figure 6.1 shows an example. The poster has tags that map to common vocabulary, e.g., author, title, abstract, etc., the content could be fully dichotomized into structured vocabulary, i.e., the information would generate separate triples describing the resource or resources referenced in the poster, or summary information could be used in the description. When shared on a server, researchers do not have to choose vocabulary because it is embedded for their research domain.

Using such templates could also help reduce issues with attribution and licensing by prompting for the information when creating the resource and then structuring with common vocabulary.

6.6 An evolving Semantic Web

Collect needs to support an evolving Semantic Web and document the source of structured representation. As the Semantic Web grows, the expectation is that more information published on the Web will have embedded microformats or self-describing URIs. When a structured representation of a resource is added to a scientific collection, **C**ollect should document the process used to structure the resource, e.g., self-described, embedded structure or RDFized. If there is a need to synchronize a scientific collection, then the process to structure a resource could change. **C**ollect should be able to adapt to the change to support and document the process. In the event that a scientific collection changes when updated,

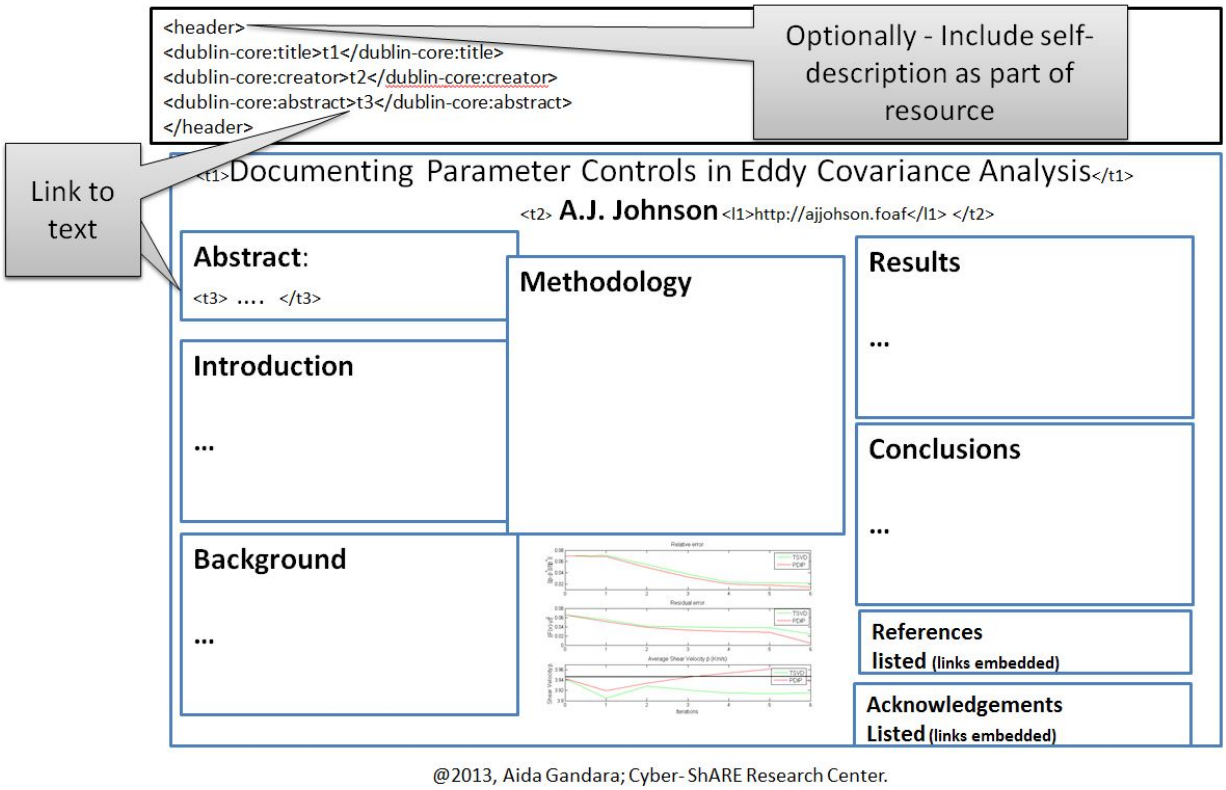

Figure 6.1: Self-describing poster template that uses default vocabulary and guides the structured documentation of research resources.

a scientific collection would then be able to justify its content. The research prototype, for example, can recreate scientific collections using the URIs that are members of the collection and URIs that are loaded in **R**efine to determine the content that will be synchronized. In cases where consistent meaning is needed, e.g., to describe scientific research, RDFizers may not be the best approach unless there is some guarantee that the reliable meaning can always be achieved. This research looks to data providers, i.e., data portals, and other Web providers to structure resources, relying on default properties when self-describing URI or embedded RDF are not available. **C**ollect can distinguish the approach and decide if information should be handled differently.

6.7 Semantic Web investment

This research determined that scientists from the case studies needed more understanding as to the benefits of the Semantic Web before they could choose vocabularies. One barrier found when documenting the case studies was a lack of semantic descriptions for the resources. There was a 2% or less Semantic Web compatibility in each of the research effort resources, meaning that there was little to no semantic descriptions created for any of the resources. Since there were no specific vocabulary requests for specific semantic vocabularies, in any of the case studies, this research concludes that obtaining scientific information on the Semantic Web will continue to be a struggle. Researchers still lack sufficient evidence that the Semantic Web can help in a scientific capacity, i.e., that they will benefit from using the Semantic Web to conduct or document research. To this end, it is important that CARP implementations produce scientific collections within existing research environments, leveraging defaults, not requiring scientists to manually choose vocabularies, and reusing existing Web content until compelling arguments emerge for scientists to customize their participation on the Semantic Web.

Chapter 7

Summary

7.1 Overview of Work

The goal of the research is to define a systematic approach for describing the resources associated with a scientific research effort such that results and related resources become more accessible and understandable to machines over the Semantic Web. The research resulted in the following: 1)definition of the CARP methodology to systematically describe a collection of research related resources as scientific collections; 2)creation of scientific collections to validate that CARP can document scientific research efforts; and 3)application of machine understandable queries to the scientific collections to verify that scientific collections are understandable to machines. The following sections present conclusions, research outcomes and future work related to the research.

The following are the outcomes of this research:

1. A methodology consisting of four phases to create scientific collections describing heterogeneous and distributed Web resources uniformly, linked and accessible over the Semantic Web.

2. A prototype system built from an existing content management system that implements the phases of the methodology. The prototype system is modular such that its parts can be adopted or replaced to enhance functionality.

3. Three case studies about three academic research efforts. The case studies elucidate the details of describing scientific research as scientific collections by documenting characteristics about the resources, the effectiveness of the methodology for each case

study and the semantic vocabulary used in each scientific collection. The case studies validate that the methodology can be applied to scientific research.

4. A set of principles that emerged from creating collections of Web-accessible resources. The principles guide the documentation process to assure scientific collections enhance the Linked Data dataspace.

7.2 Future Work

CARP has applications beyond scientific collections. For example, collections can be created to describe phenomena in a generic topic, where experts correlate Web resources and annotate them with tacit knowledge to share details that explain the phenomena. The result collection can be queried and integrated with other collections that complement or contradict the information. The remainder of this section discusses future work for CARP and scientific collections.

The CARP-based Drupal modules used to implement the prototype environment are currently in use for building *scientific collections in the study of research discovery and innovation*. The research is supported by the NSF CI-Team Virtual Learning Commons Project[1]. The VLC is investigating the contextual representation of collections to improve understanding, for example maps and tag clouds, so as to provide meaningful highlights of the resources within a research collection. The VLC's use of scientific collections are primarily leveraged for information exchange.

Most of the resources in the case studies had neither attribution or licensing set, despite over 90% of the researchers surveyed identifying both as important. One area of research for scientific collections is in understanding the application of *automated reasoning techniques on attribution and licensing* properties in a scientific collection. Questions that could be considered are: 1) what are the attribution and licensing relationships between resources

[1]VLC is supported by National Science Foundation Award number OCI-1135525 for the CI-Team Diffusion project: The Virtual Learning Commons

in a collection, 2) is an inheritance model acceptable for applying attribution or licensing rules to resources that have no setting 3) how would those rules be expressed to accurately propagate attribution and licensing expectations within a scientific collection.

Much of the work in processing resources through the CARP methodology was a manual process. The prototype provided some interfaces but there was little investment in user interfaces to capture details in the collections. For example, refine can be a tedious task of relating one URI to another, providing little direction as to why one URI is more relevant, aside from an expert identifying the need for a relationship. One area of research that would facilitate capturing research documentation are *domain-friendly abstractions* in user interfaces. For example, relationships could be created supporting a domain understanding of data and relationships such as the relationship between species. Future work for the research is to consider tools scientists use to visualize data as potential interfaces for implementing CARP to enhance scientific collections.

Currently this research is focused on understanding how to document science using Web-accessible scientific resources over the Semantic Web. The benefit of the Semantic Web is not just in the structured representation that facilitates automated query and information integration, but the ability to reason and infer new knowledge. Future work in *constraints and automated reasoning* should provide insight in the benefits that such approaches provide in enabling better analysis and more understanding of scientific collections. For example, how can the use of constraints and reasoning in scientific collections enable more understanding of scientific results and interdisciplinary research? The questions for each case study are specifically related to the properties, classes and values in the respective scientific collections. Future work is to focus on inferring new answers that are not explicitly included in a scientific collection.

The ECC Project identified privacy needs to support documenting ongoing research, i.e., there is a need for varied privacy while documenting research. Some researchers needed access to all research documentation and other collaborators required access to only some research documentation. The methodology does not address specific privacy needs, how-

ever, privacy is an ongoing discussion for scientific research (Paine et al., 2007) and Semantic Web (Artz and Gil, 2007). By default, the methodology relies on URL privacy rules of the managing system, i.e., in the case of the prototype, the Drupal content management system. In order to add new content, a user must be logged into the system. Access to URLs encodings, e.g., the SPARQL endpoint and published views, are controlled by the content management system's user-access rules as well. A separate topic from privacy of content management systems is *triple-based privacy in triplestores*.

7.3 Concluding Remarks

The CARP methodology provides the mechanisms to increase machine accessibility of research related resources in comparison to other approaches that share resources distributed and unrelated across the Web. Indeed, the main barrier to accessing result resources for the three case studies is due to the resources never being published on the Web. Even when published on the Web, many research resources are not uniquely exposed on the Web requiring manual searches. Moreover, existing searches over the Web do not facilitate locating related resources. CARP supports contextually describing collections of resources; **C**ollect identifies the members of the collection, **A**nnote and **R**efine add meaningful descriptions and relationships to describe the context of the collection and **P**ublish exposes the resources so that they can be accessed directly, through a self-describing URI. Because resources are contextually described in a scientific collection, this research demonstrates that *use of CARP allows machines to access related resources and the context around them provides the relevant information to support reuse.*

The CARP methodology provides the mechanisms to increase machine understanding of research related resources that are currently distributed and heterogeneous over the Web. The flexibility of sharing resources with different formats on the Web produces a challenge when software agents attempt to process resources of unfamiliar formats. CARP describes research related resources using a uniform representation, e.g., RDF, and exposes them as a

scientific collection accessible from a single SPARQL endpoint. In addition, the **A**nnotate and **R**efine phases add meaningful annotations and relationships to describe resources and the **P**ublish phase exposes the collection in other serializations, for use by machines or humans that can understand those serializations. By exposing meaningfully and uniformly described resources in a scientific collection, machines can query the collection to understand the meaning and relationships of the resources. Thus, *research-related resources are accessible through the context of a scientific collection that is Web-accessible and queryable, facilitating reuse of those resources*. Moreover, because scientific collections are uniformly structured and accessed through the same protocol, multiple scientific collections can be queried and processed by machines simultaneously enabling understanding across scientific research efforts as well.

The case studies confirmed that documenting scientific research on the Web is not a common practice. Since few tools are used to semantically describe scientific research results, expectations for processing distributed and heterogeneous research related information are limited to specific APIs, services and data structures. CARP guides the documentation process, by focusing on scientific details and leveraging Semantic Web tools and techniques to structure and link research resources and make them accessible on the Semantic Web. By adhering to the CARP principles, *the result is consistently a scientific collection that enhances the Linked Data dataspace*. As a result, scientific collections can leverage various Semantic Web benefits that are known challenges on the Web. For example: scientific collections can be processed aside from heterogeneous and distributed data holdings of data portals using Semantic Web techniques to analyze them (Skjveland, 2012), self-configuring data integration systems can map the resources in scientific collections to standards or data structures of other organizations (Das Sarma et al., 2008), and data mining techniques can be used to facilitate building queries about the uniformly structured research information captured in a scientific collection (Ding et al., 2006). Moreover, since scientific collections are accessible on the Linked Data dataspace, any number of tools can access, understand and process the content for other reasons.

Since CARP does not dictate either vocabulary or members of a scientific collection, many tools can participate in the documentation process. CARP functionality can be embedded in any tool that can programmatically access a scientific collection. The prototype exhibited this interoperability in the implementation of Drupal modules. Not only can modules be replaced or enhanced, but SPARQL endpoints support updates. In other words, scientific tools can embed capabilities to enhance information in a scientific collection, i.e., implement phases of CARP. Thus, *tools that implement CARP can focus on specific phases to add information to a scientific collection, relying on other tools to supplement the information.*

Scientists, in particular those funded through public organizations such as the National Science Foundation, have a responsibility when conducting research, to make findings and the justifications of those findings available to society. In addition to the NSF, organizations such as the National Institute of Health (NIH) invest millions of dollars in funding research, expecting results and findings to be reusable for future discoveries. *Scientific collections describe not just data, but all reusable resources of a scientific research effort*, including comments and relationships, from a single self-describing URI.

In conclusion, it is our goal that through the CARP methodology, scientific collections can be shared over the Semantic Web to expose findings and justifications of those findings, facilitating reuse of the results of a research effort and promoting innovative research.

References

Cyberinfrastructure for e-science. *Science*, (5723):817–821, May 2005.

CI-Server: Towards a Collective Scientific Knowledge Environment, Savannah, GA, 2010.

Documenting and Sharing Scientific Research over the Semantic Web, Graz, Austria, Sep/2012 2012. ACM.

ADIwg. Alaska data integration working group. 2013. URL `http://www.aoos.org/adiwg/`.

Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer*. MIT Press, Cambridge, MA, USA, second edition, 2008. ISBN 0262012103.

Arc2. Arc2 triplestore. 2013. URL `https://github.com/semsol/arc2/wiki`.

Donovan Artz and Yolanda Gil. A survey of trust in computer science and the semantic web. *Web Semant.*, 5:58–71, 2007. ISSN 1570-8268. doi: 10.1016/j.websem.2007.03.002. URL `http://dx.doi.org/10.1016/j.websem.2007.03.002`.

Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, New York, NY, USA, 2nd edition, 2010. ISBN 0521150116, 9780521150118.

Dave Beckett. Rdf/xml syntax specification (revised). 02/2004 2004. URL `http://www.w3.org/TR/REC-rdf-syntax/`.

Khalid Belhajjame, Oscar Corcho, Daniel Garijo, Jun Zhao, Paolo Missier, David Newman, Raul Palma, Sean Bechhofer, Esteban Garcia Cuesta, Jose Manuel Gomez-Perez, Graham Klyne, Kevin Page, Marco Roos, Jose Enriquez Ruiz, Stian Soiland-Reyes, Lourdes

Verdes-Montenegro, David C. De Roure, Carole A Goble, Frank van Harmelen, Alexander Garcia Castro, Christoph Lange, and Benjamin Good. Workflow-centric research objects: First class citizens in scholarly discourse. In *Workshop on the Semantic Publishing, 9th Extended Semantic Web Conference*, Hersonissos, Crete, Greece, May/2012 2012.

Fran ois Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, Jean Morissette, et al. Bio2rdf: towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics*, 41:706–716, 2008.

Tim Berners-Lee. Linked data, 2009. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

Tim Berners-Lee. Linked data. 2010. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, pages 34–43, May 2001.

Tim Berners-Lee, Roy Fielding, and Larry Masinter. Uniform resource identifiers (uri): generic syntax, 2005. URL `http://www.hjp.at/doc/rfc/rfc3986.html`.

Abraham Bernstein and Esther Kaufmann. Gino–a guided input natural language ontology editor. *The Semantic Web-ISWC 2006*, page 144–157, 2006.

Chris Bizer. Data integration on the public web of linked data. 2010. URL `http://www.wiwiss.fu-berlin.de/en/institute/pwo/bizer/research/publications/Bizer-FIS2010-Pay-As-You-Go-Talk.pdf`.

Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the lod cloud. 2011. URL `http://lod-cloud.net/state/`.

Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, New York, NY, USA, 2008. ACM, ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367760. URL http://doi.acm.org/10.1145/1367497.1367760.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5:1–22, Mar 2009a. ISSN 1552-6283. doi: 10.4018/jswis.2009081901. URL http://dx.doi.org/10.4018/jswis.2009081901.

Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5:22, 2009 2009b. doi: 10.4018/jswis.2009081901. URL http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901.

Shawn Bowers and Bertram Ludäscher. A calculus for propagating semantic annotations through scientific workflow queries. In *Proceedings of the 2006 international conference on Current Trends in Database Technology*, Berlin, Heidelberg, 2006. Springer-Verlag, Springer-Verlag. ISBN 3-540-46788-2, 978-3-540-46788-5. doi: 10.1007/11896548_54. URL http://dx.doi.org/10.1007/11896548_54.

Dan Brickley. Wgs84 geo positioning: an rdf vocabulary, 2006. URL http://www.w3.org/2003/01/geo/wgs84_pos$\#$.

Dan Brickley and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation. W3C, 2004. URL http://www.w3.org/TR/rdf-schema/.

Dan Brickley and Libby Miller. The Friend of a Friend (FOAF) project. 2010. URL http://xmlns.com/foaf/spec/.

Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30:107–117, 1998.

Jorge Cardoso. The semantic web vision: Where are we? *IEEE Intelligent Systems*, 22: 84–88, 2007. ISSN 1541-1672. doi: 10.1109/MIS.2007.97. URL http://dx.doi.org/10.1109/MIS.2007.97.

Chia Hui Chang, Mohammed Kayed, M.R. Girgis, and K.F. Shaalan. A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on*, 18:1411–1428, 2006. ISSN 1041-4347. doi: 10.1109/TKDE.2006.152.

citeulike. Citeulike. 2013. URL http://www.citeulike.org/.

Stéphane Corlosquet, Renaud Delbru, Tim Clark, Axel Polleres, and Stefan Decker. Produce and consume linked data with drupal! In *Proceedings of the 8th International Semantic Web Conference*, Berlin, Heidelberg, 2009. Springer-Verlag, Springer-Verlag. ISBN 978-3-642-04929-3. doi: 10.1007/978-3-642-04930-9_48. URL http://dx.doi.org/10.1007/978-3-642-04930-9_48.

Cyber-ShARE. Cyber-share projects, 2012. URL http://cybershare.utep.edu/projects.

Richard Cyganiak, Jun Zhao, Keith Alexander, and Michael Hausenblas. Vocabulary of interlinked datasets (void), 2011. URL http://rdfs.org/ns/void$\#$.

Laura Dabbish, Colleen Stuart, Jason Tsay, and Jim Herbsleb. Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, ACM, 2012.

Mathieu d'Aquin and Natalya F Noy. Where to publish and find ontologies? a survey of ontology libraries. *Web Semantics: Science, Services and Agents on the World Wide Web*, 11:96–111, 2012. doi: 10.1016/j.websem.2011.08.005. URL http://dx.doi.org/doi:10.1016/j.websem.2011.08.005.

Anish Das Sarma, Xin Dong, and Alon Halevy. Bootstrapping pay-as-you-go data integration systems. In *Proceedings of the 2008 ACM SIGMOD international conference on*

Management of data, New York, NY, USA, 2008. ACM, ACM. ISBN 978-1-60558-102-6. doi: 10.1145/1376616.1376702. URL http://doi.acm.org/10.1145/1376616.1376702.

John Davies, Alistair Duke, and York Sure. Ontoshare: a knowledge management environment for virtual communities of practice. In *Proceedings of the 2nd international conference on Knowledge capture*, New York, NY, USA, 2003. ACM, ACM. ISBN 1-58113-583-1. doi: 10.1145/945645.945652. URL http://doi.acm.org/10.1145/945645.945652.

John Day-Richter, Midori A Harris, Melissa Haendel, Suzanna Lewis, et al. Obo-edit—an ontology editor for biologists. *Bioinformatics*, 23:2198–2200, 2007.

DbPedia.org. Dbpedia. 2013. URL http://dbpedia.org.

D. De Roure, S. Bechhofer, C. Goble, and D. Newman. Scientific social objects: The social objects and multidimensional network of the myexperiment website. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, 2011. doi: 10.1109/PASSAT/SocialCom.2011.245.

David C. De Roure, Carol Anne Goble, and Robert Stevens. Designing the myexperiment virtual research environment for the social sharing of workflows. In *Proceedings of the Third IEEE International Conference on e-Science and Grid Computing*, Washington, DC, USA, 2007. IEEE Computer Society, IEEE Computer Society. ISBN 0-7695-3064-8. doi: 10.1109/E-SCIENCE.2007.29. URL http://dx.doi.org/10.1109/E-SCIENCE.2007.29.

Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: a search and metadata engine for the semantic web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, ACM, 2004.

Yihong Ding, David W. Embley, and Stephen W. Liddle. Automatic creation and simplified querying of semantic web content: an approach based on information-extraction ontologies. In *Proceedings of the First Asian conference on The Semantic Web*, Berlin, Heidelberg, 2006. Springer-Verlag, Springer-Verlag. ISBN 3-540-38329-8, 978-3-540-38329-1. doi: 10.1007/11836025_40. URL http://dx.doi.org/10.1007/11836025_40.

Drupal. Drupal, 2013. URL http://drupal.org.

Edd Dumbill. Description of a project, 2004. URL http://usefulinc.com/ns/doap$\#$.

R Fielding, J Gettys, J Mogul, H Frystyk, L Masinter, P Leach, and Tim Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1, 1999. URL http://www.hjp.at/doc/rfc/rfc2616.html.

Flickr. Flickr. 2013. URL http://www.flickr.com/.

Peter Fox and James Hendler. *Semantic e_Science: encoding meaning in the next-generation digitally enhanced science*, chapter Semantic e_Science: encoding meaning in the next-generation digitally enhanced science, pages 147–152. Microsoft Research Redmond, Washington, DC, 2009.

Michael Franklin, Alon Halevy, and David Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Rec.*, 34:27–33, 2005. ISSN 0163-5808. doi: 10.1145/1107499.1107502. URL http://doi.acm.org/10.1145/1107499.1107502.

Aída Gándara. *CI-Server Framework*. 2012. URL http://rio.cs.utep.edu/ciserver.

Aída Gándara and Hilmar Lapp. Integrating loosely structured data into the linked open data cloud, 2012. URL https://notebooks/dataone.org/lod4dataone.

Aída Gándara and Natalia Villanueva-Rosales. Documenting and sharing scientific research over the semantic web. In *Proceedings of the 12th International Conference on Knowledge*

Management and Knowledge Technologies, New York, NY, USA, 2012. ACM, ACM. ISBN 978-1-4503-1242-4. doi: 10.1145/2362456.2362480. URL http://doi.acm.org/10.1145/2362456.2362480.

Aída Gándara, George Chin, Jr., Paulo Pinheiro da Silva, Signe White, Chandrika Sivaramakrishnan, and Terence Critchlow. Knowledge annotations in scientific workflows: an implementation in kepler. In *Proceedings of the 23rd international conference on Scientific and statistical database management*, Berlin, Heidelberg, 2011a. Springer-Verlag, Springer-Verlag. ISBN 978-3-642-22350-1. URL http://dl.acm.org/citation.cfm?id=2032397.2032412.

Aída Gándara, Leonardo Salayandía, and Aline Jaimes. Ci-server framework: Cyberinfrastracture over the semantic web. In Matthew B. Jones and Corinna Gries, editors, *Proceedings of the Environmental Information Management Conference*, Portland, Or, Sep 2011b.

Paul Gearon, Alexandre Passant, and Axel Polleres. Sparql 1.1 update. 03/2013 2013. URL http://www.w3.org/TR/sparql11-update/.

Asunción Gómez-Pérez. Ontological engineering: A state of the art. *Expert Update: Knowledge Based Systems and Applied Artificial Intelligence*, 2:33–43, 1999.

Jim Gray, David T. Liu, Maria Nieto-Santisteban, Alex Szalay, David J. DeWitt, and Gerd Heber. Scientific data management in the coming decade. *SIGMOD Rec.*, 34:34–41, 2005. ISSN 0163-5808. doi: 10.1145/1107499.1107503. URL http://doi.acm.org/10.1145/1107499.1107503.

Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43:907–928, 1995. ISSN 1071-5819. doi: 10.1006/ijhc.1995.1081. URL http://dx.doi.org/10.1006/ijhc.1995.1081.

E Haber and D Oldenburg. Joint inversion: A structural approach. *Inverse problems*, 13: 63–77, 1997.

Michael Hausenblas. Utilising linked open data in applications. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, New York, NY, USA, 2011. ACM, ACM. ISBN 978-1-4503-0148-0. doi: 10.1145/1988688.1988697. URL http://doi.acm.org/10.1145/1988688.1988697.

Tom Heath. Linked Data, The Story So Far (or What Happens Next?). 2009. URL http://tomheath.com/slides/2009-09-london-linked-data-the-story-so-far.pdf.

Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1:1–136, 2011.

IRIS. Incorporated Research Institutions for Seismology (IRIS). 2012. URL www.iris.edu.

Ian Jacobs and Normal Walsh. Architecture of the world wide web, volume one. 12/2004 2004. URL http://www.w3.org/TR/webarch/.

Aline Jaimes. Understanding and scaling patterns and controls of land-atmosphere carbon, water and energy exchange in a chihuahuan desert shrubland with novel cyberinfrastructure. 09/2012 2012.

Joomla. Joomla! 2013. URL http://www.joomla.org/.

Aditya Kalyanpur, Bijan Parsia, Evren Sirin, Bernardo Cuenca Grau, and James Hendler. Swoop: A web ontology editing browser. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4:144–153, 2006.

Graham Klyne and Jeremy J. Carroll. Resource description framework (rdf): Concepts and abstract syntax. 02/2004 2004. URL http://www.w3.org/TR/rdf-concepts/.

Nick Knouf. bibtex definition in web ontology language (owl) version 0.1, 2004. URL http://zeitkunst.org/bibtex/0.2/bibtex.owl.

Marja-Riitta Koivunen and Eric Miller. W3c semantic web activity, 2001. URL http://www.w3.org/2001/12/semweb-fin/w3csw.

Peter Kraker, Derick Leony, Wolfgang Reinhardt, and Beham Günter. The case for an open science in technology enhanced learning. *Int. J. Technol. Enhanc. Learn.*, 3(6):643–654, 2011. ISSN 1753-5255. doi: 10.1504/IJTEL.2011.045454. URL http://dx.doi.org/10.1504/IJTEL.2011.045454.

Holger Lausen, Ying Ding, Michael Stollberg, Dieter Fensel, Rubén L. Hernández, and Sung-Kook Han. Semantic web portals: state-of-the-art survey. *Journal of Knowledge Management*, 9:40–49, 2005. ISSN 1367-3270. doi: 10.1108/13673270510622447. URL http://dx.doi.org/10.1108/13673270510622447.

Bo Leuf and Ward Cunningham. *The Wiki way: quick collaboration on the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. ISBN 0-201-71499-X.

LOD-Around-The-Clock. 5 star open data. 2012. URL http://5stardata.info/.

Juan Miguel López, Rosa Gil, Roberto García, Idoia Cearreta, and Nestor Garay. Towards an ontology for describing emotions. In *Proceedings of the 1st world summit on The Knowledge Society: Emerging Technologies and Information Systems for the Knowledge Society*. Springer-Verlag, Springer-Verlag, 2008. URL http://www.researchgate.net/publication/221003078_Towards_an_Ontology_for_Describing_Emotions/file/d912f507705909779c.pdf.

David Martin and John Domingue. Semantic web services, part 2. *Intelligent Systems, IEEE*, 22:8–15, 2007.

Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. 2003. URL http://www.w3.org/TR/owl-features/.

Alistair Miles and Sean Bechhofer. Simple knowledge organization system, 2004. URL http://www.w3.org/2004/02/skos/core$\#$.

Natalya F Noy. Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, 33:65, 2004.

Natalya F Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W Fergerson, and Mark A Musen. Creating semantic web contents with protege-2000. *Intelligent Systems, IEEE*, 16:60–71, 2001.

Ontotext AD. Owlim. 2013. URL `http://www.ontotext.com/owlim`.

Openlink Software. Virtuoso universal server. 2013. URL `http://virtuoso.openlinksw.com/`.

Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello. Sindice.com: a document oriented lookup index for open linked data. *International Journal of Metadata and Semantics and Ontologies*, 3:37–52, 2008. ISSN 1744-2621. doi: 10.1504/IJMSO.2008.021204. URL `http://dx.doi.org/10.1504/IJMSO.2008.021204`.

ORNL. Ornl-daac oak ridge national lab distributed active archive center. 2012. URL `http://daac.ornl.gov/`.

Carina Paine, Ulf-Dietrich Reips, Stefan Stieger, Adam Joinson, and Tom Buchanan. Internet users' perceptions of 'privacy concerns' and 'privacy actions'. *Int. J. Hum.-Comput. Stud.*, 65:526–536, 2007. ISSN 1071-5819. doi: 10.1016/j.ijhcs.2006.12.001. URL `http://dx.doi.org/10.1016/j.ijhcs.2006.12.001`.

Nick Pearce, Martin Weller, Eileen Scanlon, and Sam Kinsley. Digital Scholarship Considered: How New Technologies Could Transform Academic Work. *In Education*, 16, 2010.

Paulo Pinheiro da Silva, Leonardo Salayandía, Nicholas Del Rio, and Ann Q. Gates. On the Use of Abstract Workflows to Capture Scientific Process Provenance. In *Proceedings*

of the 2nd Workshop on the Theory and Pratice of Provenance (TaPP'10) at USENIX, San Jose, CA, 2010.

María Poveda-Villalón, Mari Carmen Suárez-Figueroa, and Asunción Gómez-Pérez. Validating ontologies with oops! In *Proceedings of the 18th international conference on Knowledge Engineering and Knowledge Management*, Berlin, Heidelberg, 2012. Springer-Verlag, Springer-Verlag. ISBN 978-3-642-33875-5. doi: 10.1007/978-3-642-33876-2_24. URL `http://dx.doi.org/10.1007/978-3-642-33876-2_24`.

Eric Prud'hommeaux and Andy Seaborne. Sparql query language for rdf, 15 January 2008.

PublishMyData. Publishmydata. 2013. URL `http://publishmydata.com/`.

Bastian Quilitz and Ulf Leser. *Querying distributed RDF data sources with SPARQL*, page 524–538. Springer, 2008.

readcube. readcube. 2013. URL `http://www.readcube.com/`.

OJ Reichman, Matthew B. Jones, and Mark P Schildhauer. Challenges and opportunities of open data in ecology. *Science(Washington)*, 331:703–705, 2011.

Richard Reid and Peter Edwards. Ourspaces-a social semantic web environment for escience. *Proceedings of the AAAI2009*, 2009.

Leo Sauermann and Richard Cyganiak. Cool uris for the semantic web. 12/2008 2008. URL `http://www.w3.org/TR/cooluris/`.

SemanticWeb.org. Tools. 2013. URL `http://semanticweb.org/wiki/Tools`.

Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21:96–101, 2006. ISSN 1541-1672. doi: http://doi.ieeecomputersociety.org/10.1109/MIS.2006.62.

Ben Shneiderman. Computer science: Science 2.0. *Science*, 319(5868):1349–1350, March 2008.

Martin G. Skjveland. Sgvizler: A javascript wrapper for easy visualization of sparql result sets. In *9th Extended Semantic Web Conference*, May/2012 2012.

Uram Anibal Sosa. On constrained optimization schemes for geophysical inversion of seismic data. Dissertation, 11/2012 2012.

Ahmet Soylu and Patrick De Causmaecker. Embedded semantics empowering context-aware pervasive computing environments. In *Proceedings of the 2009 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, Washington, DC, USA, 2009. IEEE Computer Society, IEEE Computer Society. ISBN 978-0-7695-3737-5. doi: 10.1109/UIC-ATC.2009.12. URL `http://dx.doi.org/10.1109/UIC-ATC.2009.12`.

Carly Strasser, Robert Cook, William Michener, Amber Budden, and Rebecca Koskela. Dataone promoting data stewardship through best practices. In Matthew B. Jones and Corinna Gries, editors, *Proceedings of the Environmental Information Management Conference*, Santa Barbara, CA, Jul 2011.

York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann, and Daniel Oberle. *The SWRC ontology–Semantic Web for research communities*, page 218–231. Springer, 2005.

The Knowledge Network for Biocomplexity. The knowledge network for biocomplexity (knb). 2012. URL `http://knb.ecoinformatics.org/index.jsp/`.

Lennox E Thompson. Seismic investigation of the southern rio grande rift. Master's thesis, 04/2010 2010.

Herbert van Sompel and Carl Lagoze. *All Aboard: Toward Machine-Friendly Scholarly Communication System*, pages 147–152. Microsoft Research, Washington, DC, 2009.

W3C. W3c semantic web activity. 2012. URL `http://www.w3.org/2001/sw`.

Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413:222, 1998. URL `http://dl.acm.org/citation.cfm?id=RFC2413`.

Wordpress. Wordpress. 2013. URL http://en.wordpress.com/about/.

Zotero. Zotero. 2013. URL http://www.zotero.org/.

Glossary

content management system a Web server that manages information that will be available on the Web. A content management system is normally supported with a back-end database and directory structure to aid in holding Web content. In addition, a content management system is exposed through an HTML menu driven interface that supports user-based privacy of Web content.

Linked Data A Semantic Web approach where descriptions of include properties that link to other resources. In this way, the data (resources) are linked

linked open data A linked data approach where (resources) are linked openly without the need to enter passwords.

pay-as-you-go a test

resource Any digital object that is accessible over the Web. A resource can be identified by a URI(Jacobs and Walsh, 2004)

self-describing URI A Web resource that has been enabled to describe itself, i.e., its contents, when the resources URI is accessed.

Semantic Web Interconnected information that describes the meaning of information on the Web. A web client, e.g., a semantic browser, can request a description of a Web resource instead of the content itself.

SPARQL A query language and protocol for querying RDF from a database. SPARQL stands for SPARQL Protocol and RDF Query Language

SPARQL endpoint A URL that has been enabled to accept SPARQL queries and return results as RDF triples.

triple a test

URI a unique Web address

URL a unique Web address that results in a HTML-based Web page

Acronyms

ADIwg Alaska Data Integration Working Group

CARP **C**ollect-**A**nnotate-**R**efine-**P**ublish Methodology

CMS Content Management System

DataONE Data Observation Network for Earth

DOAP Description of a Project

DOIs Digital Object Identifiers

EML Ecological Metadata Language

HTML Hypertext Markup Language

HTTP Hypertext Transfer Protocol

KNB Knowledge Network for Biocomplexity

LOD Cloud Linked Open Data Cloud

NIH National Institute of Health

NSF National Science Foundation

OBFS Organization of Biological Field Stations

OWL the Web Ontology Language

RDF Resource Description Framework

RDF-S Resource Description Framework Schema

SKOS Simple Knowledge Organization System

voiD Vocabulary of Interlinked Datasets

Web World Wide Web

XML Extensible Markup Language

Appendix A

The Initial Methodology

The initial work with documenting research was focused on a Web platform, CI-Server, for identifying related research results that were accessible on the Web. For each project we automatically built pages to show different perspectives of the work including description, resources and searches. The resource page listed each resource, categorized by type, with a link to a tool that could open and view each resource. This was done for collaborators and team members of the Cyber-ShARE Trust research team(Gándara et al., 2011b).

In terms of accessibility, the approach had some small but important successes and findings for this research. First, the platform, through services and a client API, enabled software agents to publish work on the server that were usually in standalone systems. We had observed how the use of local files was resulting in duplicates of the same file, confusion on the latest version and the need to manually reset internal links when a file was uploaded to a Web server. Once the Trust tools were instrumented with these Web sharing capabilities, sharing on the Web was the preferred choice, increasing the files on the server to thousands within a few months. Second, documenting research exposed the importance of linking to things to avoid duplication and of supporting some level of attribution. This was particularly seen in cases with duplicate files on different systems. In addition, we noticed that many of the tools created by the Trust team solved specific problems, e.g., capturing provenance, describing work processes, querying semantic information, but they were disconnected from each other because they were limited in how they could share information. For example, a query tool accessed provenance published on the server by extracting content published by a provenance tool on the CI-Server Webpages, the two tools had no direct way to exchange information. Several issues arose such as different

URIs for the same resource and some resources not showing up.

From discussions with the group and other collaborators, we concluded that the collections were a positive step in sharing research results and in understanding research simply because they were published in the context of each other. However, there was still a challenge because really understanding things required users to open up different files and compare things. There was a need to automate some of the manual steps, in particular searches, e.g., searching for properties of heterogeneous files, hence the decision to leverage Semantic Web techniques.

The initial methodology for transforming the collections into a machine understandable research effort is shown in Figure A.1(Gándara and Villanueva-Rosales, 2012). Basically the diagram suggests that management of research related resources go through a cycle of four states, COLLECT, where resources are identified and added to the collection, COLLABORATE: where annotations and comments are collected about resources, TRANSFORM: where the resources are transformed to an ontology representing the semantic research description and PUBLISH: where the resources are processed for publishing on the Web. Publishing a research effort is by default a Web page with the list of resources. The resources are displayed using HTML when dereferencing for user views and an ontology when dereferencing to a semantic representation.

There are a few drawbacks to this initial methodology. First, once a resource is collected, it is not collected again, so a cyclic representation of the different phases was misleading. What actually happens is that a resource is collected and then it can be annotated and transformed and shared on the Web as needed. The second issue is that using the term collaboration for the second phase might imply an interest in modeling collaborative environments, for example, chat tools, wikis and blogs, which we are not. The real purpose for the COLLABORATE phase is to capture the relationships between unstructured information such as text to research resources, explicitly. In addition, we noticed we needed a phase to harvest semantic descriptions about a resource already on the Web and to relate new or existing resources to other members of a research collection. Such refinements would

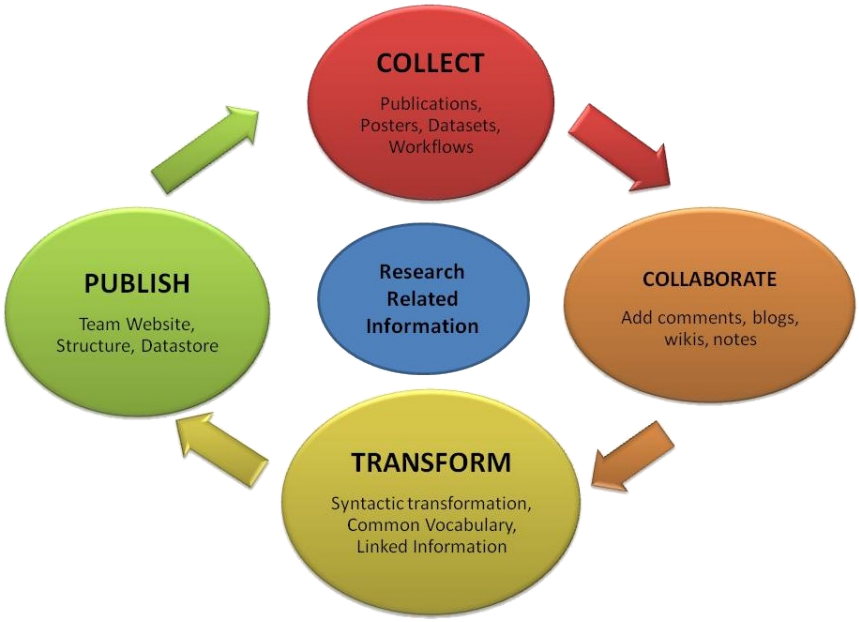

Figure A.1: The initial methodology

be necessary to explicitly expose relationships about research resources, e.g., if a poster describes an output file. The final issue was that deferring the transformation of research resources to later in the methodology meant that in the meantime, the system was working with a server-based identification of each resource.

This initial approach did not require support of a triplestore. When considering pulling more information about a resource into the system or identifying a relationship between resources, the preference is to work with a URI instead of assigning some temporary value until the transformation occurred. It was decided that the transformation should occur during COLLECT at which point all other phases of accessing and working with resources would reference a Web-accessible URI.

Appendix B

An Initial User Survey

An initial survey was conducted with one research team (9 team members) to assess each researchers opinions concerning research resource sharing. This survey was conducted prior to conducting the three case studies. Initially, each researcher was asked to provide the types of resources used or created in their research. This list is described below. In addition, each researcher was asked to answer 9 survey questions. This appendix will briefly present the information that was collected.

Resource Types

The following resource types were identified by the researchers in this survey. In addition, the case studies managed the same list of resource types.

- image and animation - individual files that contain a picture or an animation. Pictures were not extracted from inside of posters, presentations or publications.
- workflow - a diagram depicting a process used within the research effort.
- program - software program code used to conduct research. This included scripts and programs, usually in matlab, created by a scientist or other collaborators. Some of these programs worked with other tools, e.g., the CO Project's PDIP program was integrated with a joint inversion tool, which were either downloaded to their systems.
- tool - software programs used to conduct research. All three of the research projects used Software tools like Matlab to write programs or visualization tools to analyze the results of their data. Some of the tools were downloaded to their system and

integrated with the directory they were working on and others were online tools where data was uploaded and results were seen or downloaded.

- presentation - slides created by a researcher of the research effort, that describes a topic about the research. Some of these were presentations at conferences, others were created for presentations at The University.
- poster - a single diagram describing the research, each having an introduction, methodology and results.
- publication - a written document explaining some topic within the research. In some cases these were peer reviewed publications, in other cases they were administrative a, e.g., instructions for replacing sensor cards.
- data - a file or archive containing data, such as a table or parameters.

These resource types are referenced in Tables 5.1, 5.4 and 5.7.

Survey Questions

The first three questions are concerned with accessibility of resources, asking how important each researcher feels accessibility of resources is in reusing research resources; how accessible a researcher feels their resources are; and if a researcher feels they need to increase accessibility of the resources they create. Figure B.1 shows researchers agree that accessibility is important for reuse, the majority feels that their resources are accessible yet they unanimously agree (or strongly agree) that they need to increase accessibility.

The next three questions ask about the importance of a resource being understandable for reuse of research resources. The first question asks if a researcher feels that it is important that resources be understandable to reuse resources; the second asks if a researcher believes their resources are understandable and the third asks if they think they need to increase understanding of resources. Figure B.2 shows researchers agree that understanding

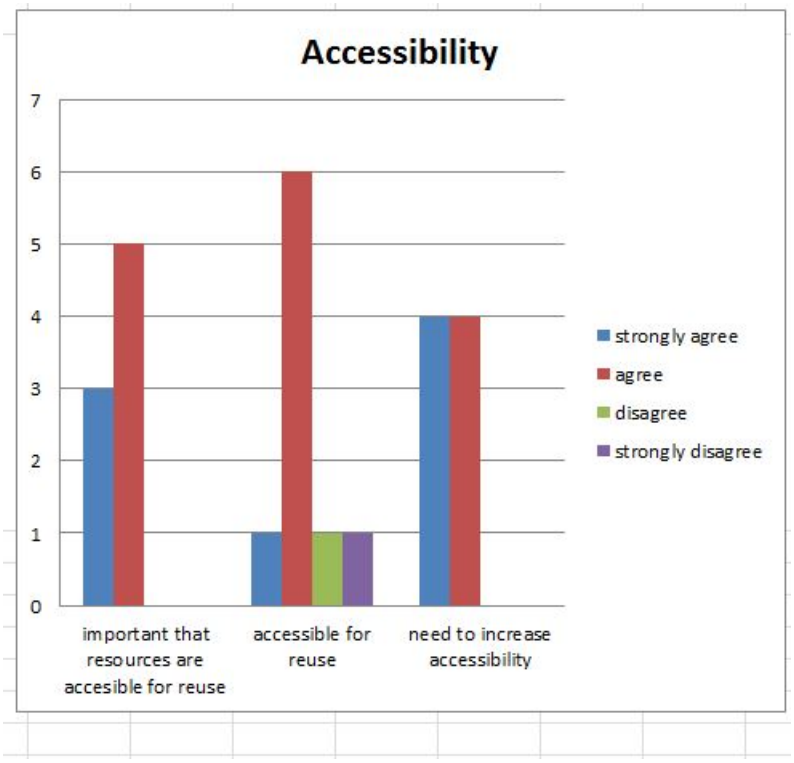

Figure B.1: Results for the initial survey on accessibility

is important to reusing resources, their answers are mixed on whether they believe their resources are understandable and unanimous concerning the need to increase understanding of resources.

The final three questions are concerned with attribution, i.e., how important researchers feel attribution is for the reuse of research resources. The first question asks if a researcher feels attribution is important; the second question asks if a researcher feels their resources have attribution (to themselves); and the third question asks if researchers feel they need to increase attribution. Figure B.3 shows that researchers agree that attribution is important to resource reuse, most of the researchers agree or strongly agree that the resources they create have attribution, yet the majority strongly agrees that they need to improve attribution.

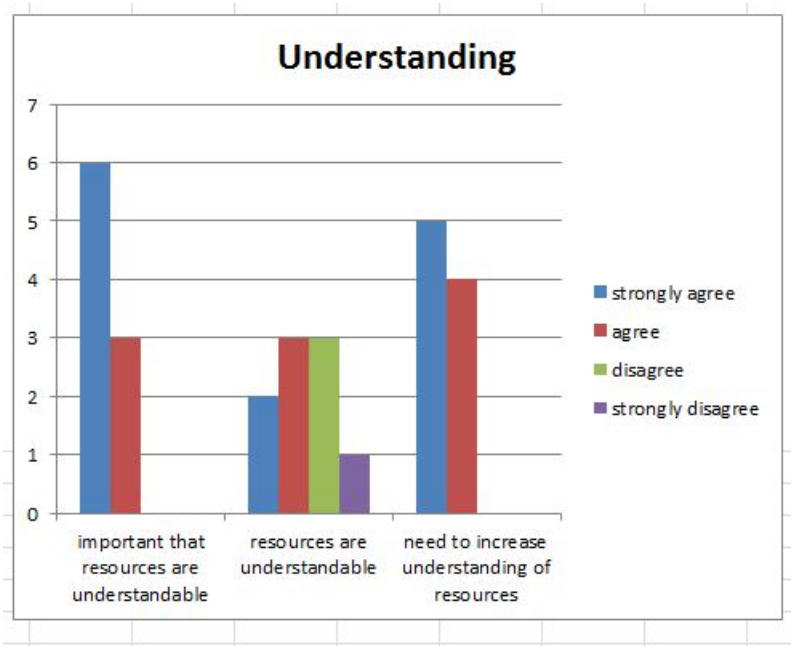

Figure B.2: Results for the initial survey on understanding of resources

Through this survey an initial resource list was collected and, in fact, was similar to the list of resources created and used by Cyber-ShARE researchers in the case studies. In addition, since the goal of this research is to increase understanding and accessibility of research resources, the survey was expected to obtain a small sampling assuring researchers are interested in those topics as well.

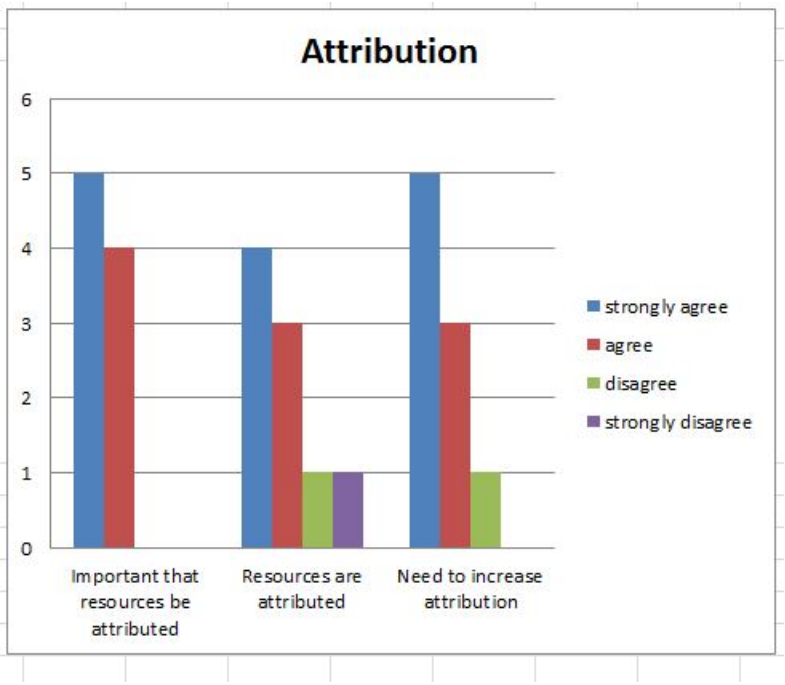

Figure B.3: Results for the initial survey on attribution of resources

Curriculum Vitae

Aída Gándara earned her Bachelor and Master of Computer Science degrees at The University of Texas at El Paso (UTEP) in 1990 and 1994, respectively. For the next 13 years, she started a family, was a lecturer for the Department of Computer Science at UTEP, worked in industry in El Paso, TX and Palo Ato, CA, and participated in efforts in her community to expose and train people to use new technologies. In 2007, she joined the doctoral program at The University of Texas at El Paso.

Ms. Gándara has been the recipient of numerous honors and awards. She is an AGEP Scholar, a CREST Scholar and a WISE Scholar, through which she had the opportunity to meet other scientists and receive guidance on her academic career. She was also the recipient of The Patricia and Jonathan Rogers Scholarship in Graduate Engineering. While pursuing her degree, Ms. Gándara worked as a research associate for the Cyber-ShARE Research Center of Excellence. Ms. Gándara collaborated with members of the Pacific Northwest National Laboratorys SciDAC Group in studying documentation of collaborative scientific research using scientific workflows and she participated as a DataONE summer intern studying Linked Data principles applied to scientific data.

Ms. Gándara has published her research at conferences in the US and internationally, in both computing and scientific communities. Her most recent publication titled "Documenting and Sharing Scientific Research Over the Semantic Web" was published in the 12th International Conference on Knowledge Management and Knowledge Technologies (2012), in Graz, Austria. Ms. Gándara's dissertation entitled, "A Semantic Web-based Methodology For Describing Scientific Research Efforts", was supervised by Dr. Natalia Villanueva-Rosales and Dr. Ann Quiroz Gates. Ms. Gándara completed her PhD in August of 2013.

Permanent address: 1116 Cerrito Bajo Lane
El Paso, Texas 79912-3555