Table of Contents

0 Introduction 2

1 Scientific information infrastructure 4

1.1 From infrastructure to information infrastructure 4

1.2 From information infrastructure to data-driven information infrastructure 5

1.3 Requirements for cyberinfrastructure 7

1.3.1 Reproducibility 7

1.3.2 Bigness of data 7

1.3.3 Data objects as the end research outputs 9

1.3.4 Open science 9

1.3.5 Distributing science 10

1.4 Elements of cyberinfrastructure 11

1.4.1 Identification system 11

1.4.2 Metadata standards 12

1.4.3 Data paper and data journal 13

1.4.4 Virtual research environment 14

2 Data and software in the research lifecycle 16

2.1 Data and its epistemologies 16

2.1.1 Data as raw and unprocessed materials 17

2.1.2 Data as means of scientific studies 18

2.2 Scientific software 19

2.2.1 Definition of software concepts 19

2.2.2 Scientific software and its classification 19

2.3 Laboratory studies 20

2.3.1 Experimental system 21

2.3.2 Norms of scientific writing and scholarly communication 22

2.4 The complex of data-software-method-theory 24

2.4.1 Package of scientific works and the theory-method package 24

2.4.2 Data and software 26

2.4.3 Data objects and theoretical objects 27

3 Citation theories and their applications on data objects 29

3.1 Three “paradigms” of citation theory 29

3.1.1 The normative theory 30

3.1.2 The rhetoric theory 32

3.1.3 The symbolic theory 33

3.2 Relationship between the three models and their applicability to the studies of data objects 35

3.3 Content and context analysis and its descendant 36

3.4 From documents to entities 37

4 Conclusion 40

REFERENCE 41

0 Introduction

This literature review is about data and software in data-driven scientific activities: the roles they play separately and collectively, and the relationship between data, software, and other scientific objects during the production of knowledge. The importance of this question lies in the fact that data and software have been holding increasingly more eminent positions in scientific studies during the past decade. Their renowned status in science did not become a reality until the rise of the new scientific paradigm, or the "fourth paradigm" (Hey, Tansley, Tolle, & others, 2009). Thanks to the growing technical capacities, both data and software are so fundamental in the contemporary scientific practices that they are introducing new epistemologies of scientific studies and new requirements for the various components of the scientific infrastructure as well as scientists themselves. What is equally important but less reported is that they are also born from, and thus inevitably shaped by the existing scientific infrastructure in extraordinary ways, which is another aspect of their deep entanglement with other scientific/extra-scientific objects.

Chapter one of this review summarizes the data-driven space of scientific information (or "cyberinfrastructure") as the background of the proposed study. The concepts and history of cyberinfrastructure are examined. What follows is a discussion of some higher-level characteristics for cyberinfrastructure, including reproducibility, bigness of data, data objects as the end research outputs, openness and distributedness of science. At the end of this chapter, some key components of cyberinfrastructure are reviewed, including identification systems, metadata standards, data papers and data journals, and virtual research environments.

Data and software that are used in scientific studies are naturally scientific objects, which is the topic of the rest of the document. More specifically, we will discuss how they facilitate the production of scientific knowledge, and how the various relations formed around data and software as they participate in this process. Given the importance of these objects, they have been studied in various knowledge domains, including but not limited to science and technology studies (STS) and information science. In these two knowledge domains, two specific research paths are especially relevant to this proposed study, namely laboratory studies and studies of citation theory.

These two areas are selected because each of them owns a set of theories and methods that can describe important aspects the topic of this literature review. These two theory-method packages (see the discussion in section 2.3.1) are distinct but related. Their connectedness is not without historical roots that can be traced back to the beginning of each field. But as they have been growing apart since the late 1970s, we are standing at a point where such connectedness can and should be reestablished to form a more comprehensive understanding of scientific activities under cyberinfrastructure.

Laboratory studies offer direct observations of how scientific knowledge is produced in the laboratories. By "*in situ* monitoring of contemporaneous scientific activity" (Woolgar, 1982, p. 484), this line of research focuses on both the actual and dynamic processes of science and how these processes are shaped by different human and non-human participants. Thus, besides a more accurate description of the drafted nature of the production of scientific knowledge, laboratory studies also give us the chance to better understand how knowledge production is interwoven with both the social and technical textures of the scientific systems.

This perspective of data and software is the main topic of Chapter two. This chapter begins with a review of the definitions and epistemological stances around data and software. What follows is how knowledge production is defined in laboratory studies, i.e., scientific knowledge is produced as scientific objects "flow" from laboratory space to scientific publication. These two spaces are the main locations where scientific objects are transformed, translated and interpreted into scientific knowledge. As these objects move between these two spaces, they are shaped by the different textures and requirements of both spaces. As data and software go through these transformations, they are frequently merged into "packages" along with other related entities, such as scientific theories, methods, and models. The concept of package is derived from STS and has significant implications towards both how knowledge is created and communicated. How package is discussed by STS researchers and some potential packages formed by data objects are discussed in this chapter. By including all these discussions, this chapter forms a theoretical framework of not only a biography or journey of scientific data objects (Daston, 2000; Leonelli, 2014), but also how their lifecycles intersect with the lifecycles of experiments as well as scientific writing, a question that is fundamental in both STS and information science.

Located in information science, studies of citation theory aim at understanding the behaviors and patterns of bibliographic citations as well as the relationship between citing and cited documents and objects, based on the documents where these citations are inscribed. Even though both givers and receivers of the citation were scientific papers in the past, as they become more important, data objects are increasingly visible in the network of citation. Thus, citation analysis can help us understand how data objects are involved in scientific studies and writing from a more textual perspective.

Chapter three reviews three specific paradigms of citation theories, namely the normative theory, the rhetoric theory, and the symbolic theory. For each paradigm, how it uniquely perceives the meanings of citation and the inter-document(object) relationship is specifically addressed in our discussion. Moreover, this chapter also reviews content and context analysis, a method that is connected to all three paradigms and is especially useful for understanding the contexts of bibliographic citations. At the end of this chapter, the scholarship of digital object citation is also reviewed, with their connections to both the citation theories and digital scholarship discussed above.

1 Scientific information infrastructure

This chapter discusses the environments in which data-driven scientific activities are operated. All scientific works happen in some sorts of infrastructure – the preceding works to support what is to follow. Given its importance to science, a brief history and the concepts of cyberinfrastructure are reviewed in the beginning of this chapter. After these introductions, some higher-level requirements for cyberinfrastructure are extracted from literature and discussed. Most, if not all, of these requirements are also those to be fulfilled by data and software objects. Before the end of this chapter, some key components of cyberinfrastructure that are related to data and software are discussed, including identification systems, metadata standards, data papers and journals, and virtual research environments, which set up the spaces where data and software will be examined in future studies.

1.1 From infrastructure to information infrastructure

An infrastructure is an aggregated "prior work that supports and enables the activity we are … engaged in doing" (Slota & Bowker, 2016). It has been an important topic in both science and technology studies and information science during the past decades, because of the notion that infrastructures hold values as they accommodate and block certain kinds of activities and relations. This notion is reflected in Langdon Winner's famous discussion of "Do Artifacts Have Politics?" (Winner, 1980), where the author discussed how the roads and bridges designed by Robert Moses followed certain specifications so that their use by "lower" classes was discouraged.

It is a consensus among researchers that infrastructures are relational, rather than substantial. First, being highly heterogeneous, infrastructures are beyond any single physical equipment or abstract entity such as protocols, standards, and memory (Bowker, Baker, Millerand, & Ribes, 2009). Second, any infrastructure has to be "sunk" in other technological and social infrastructures (Star & Ruhleder, 1994). For example, cars are useless without the infrastructures of modern transportation and energy, which themselves are reliant upon other pieces of infrastructure. But more importantly, being relational also suggests that infrastructures are much more about "when" than "what". An often-cited example is Engeström's discussion of "When is a tool" (1990); in this paper, Engeström argued that a tool is not defined by pre-given attributes: a tool becomes a tool in particular activities. On the same page, a piece of infrastructure in one context could be the working object in another context (Star & Ruhleder, 1994).

An important class of infrastructures deals with information. They are normally called information infrastructure or knowledge infrastructure. Both terms share a vagueness of definition, mirroring the difficulties to define the concepts of data, information, and knowledge. What seems to be agreed by many researchers is that information infrastructure is a broader term than knowledge infrastructure, the latter of which is more focused on organizing and processing information (Edwards, 2010; Harvard Information Infrastructure Project, 1995; Kahin, 1993). At the same time, several other researchers simply use these two terms together without distinguishing them (e.g., Lambe, 2014;

Maes, Rijsenbrij, Truijens, Goedvolk, & others, 2000). Across this literature review, we will largely use these two concepts interchangeably to suggest all the potential elements of scientific works related to scientific data and software. In the following sections, we will discuss the recent development of a data-driven paradigm of scientific practice and infrastructure, or "the fourth paradigm" (Hey et al., 2009).

1.2 From information infrastructure to data-driven information infrastructure

The variety of concepts to describe the idea of data-driven studies is a palpable phenomenon. These concepts include cyberscience (Nentwich, 2003; Nentwich & König, 2012), e-Science (Atkins, 2003; Hey & Trefethen, 2002), cyberinfrastructure (Atkins, 2003), digital scholarship (Unsworth, 2006), and data scholarship (Borgman, 2015), just to name a few commonly-used ones. The last two terms are preferred in this literature review because they are broad enough to cover multiple scenarios in which data and scientific studies are overlapped. For example, digital scholarship includes the following tasks from collecting data to creating tools to using these data objects in scientific studies (Unsworth, 2006):

1. Building a digital collection of information for further study and analysis
2. Creating appropriate tools for collection-building
3. Creating appropriate tools for the analysis and study of collections
4. Using digital collections and analytical tools to generate new intellectual products
5. Creating authoring tools for these new intellectual products, either in traditional forms or in digital form (p. 7)

Besides the semantic differences between these terms, they also form a chronology of the general idea of combining research, data, and technologies together. "Cyberscience" is one of the earliest terms researchers used to represent this complex of concepts. Figure 1.1 shows the frequencies of the concepts of "cyberscience", "e-Science" and "cyberinfrastructure" shown in the Google Books Ngram Viewer in April 17, 2017. The results indicate that cyberscience was used as early as early-1990s, while both e-Science and cyberinfrastructure were not used in any publication published before 2000 that is indexed by the Google Books Project.

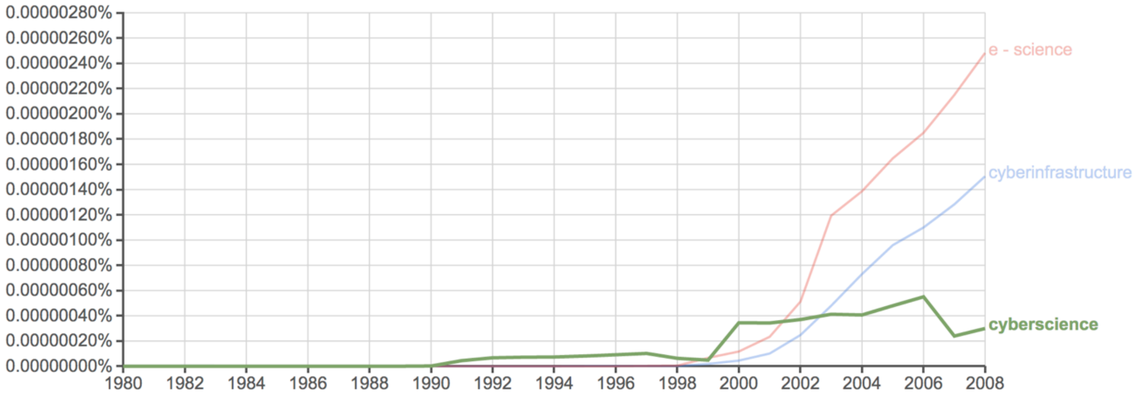


Figure 1.1: Frequencies of "e-Science", "cyberinfrastructure", and "cyberscience" in Google Books Project

The earliest instance of using "cyberscience" in English is Evelyn Fox Keller's book *Refiguring Life* (Keller, 1995). Keller (1995) coined this term to refer to the efforts to combine biology with "information theory, cybernetics, systems analysis, operations research, and computer science" (p. 86), which is a new reality in biology where scientific practice and modern information technologies are merged together.

Even though Keller's work is not included in the reviews conducted by Nentwich and his colleagues (Nentwich, 2003; Nentwich & König, 2012), this concept was given a similar yet more comprehensive definition by Nentwich later:

> "all scholarly and scientific research activities in the virtual space generated by the networked computers and by advanced information and communication technologies, in general." (Nentwich, 2003, p. 22)

As relatively later comers, e-Science and cyberinfrastructure eventually gained stronger popularities than cyberscience, mostly thanks to the famous report finished by Daniel Atkins on behalf of the National Science Foundation (Atkins, 2003). In this report, he discussed the e-Science Core Programme that was initiated in UK in the early 2000s (Hey & Trefethen, 2002) and proposed an advanced cyberinfrastructure program in the US. Some of the key definitions of these two terms are summarized in Table 1.

Concept	Definition
e-Science	"e-Science is about global collaboration in key areas of science and the next generation of infrastructure and will enable it" (Hey & Trefethen, 2002, p. 1017)
e-Science	"the application of computer technology to the undertaking of modern scientific investigation, including the preparation, experimentation, data collection, results dissemination, and long-term storage and accessibility of all materials generated through the scientific process" (Bohle, 2013)
cyberinfrastructure	"infrastructure based upon distributed computer, information and communication technology" (Atkins, 2003, p. 5).

Table 1: Summary of concepts related to data scholarship

Based on the definitions above, it is obvious that e-Science and cyberinfrastructure bear similar meanings with cyberscience, in the sense that they are the combination of scientific research and modern information technologies. About the relationship between these two concepts, a popular view is that they are synonyms used in different geographic regions (Gold, 2007; Sahoo, Sheth, & Henson, 2008). Some researchers pointed out that e-Science and cyberinfrastructure are on separate levels; the corresponding term to cyberinfrastructure should be e-Science application (Lynch, 2006; A. Zimmerman, 2007).

No matter which view we agree, e-Science and cyberinfrastructure (and other synonyms) together depict a new sociotechnical space where data and computational capacities are plugged into the traditional research world. This new world is faced with old and new requirements, and is composed of old and new components. These two topics will be addressed in the following two sections, respectively.

1.3 Requirements for cyberinfrastructure

As a sociotechnical space that pre-exists and supports the new paradigm of scientific studies, cyberinfrastructure is bound by certain requirements to finish all the necessary tasks it promises. In this section, we review some of the most important requirements in the literature, namely reproducibility, bigness of data, data objects as the end research outputs, open science, and science as a distributed endeavor. For each item, how it is posited between the old and new scientific paradigms and how it is related to the topic of this literature review are especially emphasized.

1.3.1 Reproducibility

Reproducible research means that all researchers can recreate the full analyses described in the publication (Gentleman & Lang, 2004). Strongly connected to computational science, this requirement specifies that "all data, software source code, and tools" (Qin, Dobreski, & Brown, 2016, p. 220) should be distributed along with the publication itself. Replication has been one of the most basic principles in science (National Research Council, 2011). But it becomes more important as more studies take an computational approach and scientists are equipped with stronger computational capacities (Gentleman, 2005; Ince, Hatton, & Graham-Cumming, 2012; King, 1995; Laine, Goodman, Griswold, & Sox, 2007; Peng, 2011). Often underlying these calls is the fact that many studies fail to be reproduced in following-up experiments, a phenomenon that is called "reproducibility crisis" (Baker, 2016; Peng, 2015).

In short, the requirement of reproducibility marks the trend in which science is increasingly evaluated by the processes rather than the final products (Leonelli, 2014). As two key players along the (re-)production of scientific knowledge, datasets and software are increasingly valued in this regard. First, it is important if not imperative to identify data and software entities used in scientific studies, with the help of many components of cyberinfrastructure that will be reviewed in the coming section. Second, it is also more critical to connect data objects with each other as well as with research outputs, which is an important context of this proposed study.

1.3.2 Bigness of data

As a concept that is popular in both academic and public discourses, big data is used in many cases to denote to the general data-driven scientific paradigm. Popularities of this concept accompany the vast number of definitions big data has received. After reviewing these definitions, Ward and Barker (2013) concluded that all definitions adopt at least one of the following criteria, namely size, complexity, and technologies. For example, Doug

Laney's famous definition (2001) touches three characteristics of big data: volume, velocity, and variety. Kitchin (2014) pointed out that it is the share of all the three attributes proposed by Laney that distinguishes big data from traditional databases or other types of media.

Among other characteristics, bigness is the most closely connected to big data, even though many researchers have pointed out the difficulties to define bigness (Borgman, 2015; Manovich, 2011). To address this difficulty, some researchers resort to relative quantity (Gandomi & Haider, 2015; Ward & Barker, 2013). One scale with which bigness is compared is the technical capacity. Boyd and Crawford (2012) observed that big data "is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (p. 663).

In this regard, big data is obviously facilitated by the existing information infrastructure, but at the same time challenges many aspects of scholarship. In his powerful declaration of big data, Chris Anderson (2008) announced:

> "At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right."

As data becomes bigger, it becomes "commodity and recognised output" (p. 1) in science, which further brings new requirements on the methods, infrastructures, technologies, and skills to handle the data (Leonelli, 2014). What comes after, claimed by popular narratives, is the triumph of pragmatic mathematics over scientific theories (Mayer-Schönberger & Cukier, 2013), correlation over causation (Anderson, 2008), and inductive reasoning over deductive reasoning (Erevelles, Fukawa, & Swayne, 2016; Frické, 2015; Krumholz, 2014).

However, not everyone agrees with these narratives. For example, objections have been raised to the ideas that theories can be separated from data and statistics, that big data does not need to rely upon pre-existing theories and post hoc human interpretation, that big data is of higher quality, as well as its completeness and privacy issues (Boyd & Crawford, 2012; Frické, 2015; Leonelli, 2014; Manovich, 2011; Norvig, 2008; Van Dijck, 2014).

As a (over-)hyped term, it is undeniable that the level of bigness of data is significantly different in various knowledge domains and even research communities. The strongest implication of bigness of data for this literature review is that as a field manages and analyzes a larger amount of data, if these data objects play different roles in the creation of scientific knowledge, and if they are having different relationship with other scientific entities as we will discuss in the next chapter? If so, how these differences are reflected in scientific representations?

1.3.3 Data objects as the end research outputs

It is commonly claimed that data is becoming a "first-class research object," which means that data objects are valued for the sake of themselves (Burton & Treloar, 2009; Force & Robinson, 2014). This is a dramatic change from a traditional research norm, where data would be no longer the focus in research activities after the results are published (Borgman et al., 2013).

The requirement that data objects should be the end research output cannot be better reflected by the fact that publishing datasets becomes a creditable academic activity. This activity is facilitated by the policies developed by research associations and institutions (Field et al., 2009; National Institutes of Health, 2003; Schofield et al., 2009) and journals (Baker, 2012; Bloom, Ganley, & Winker, 2014; Piwowar & Chapman, 2008), as well as new forms of scholarly communication such as data papers and data journals. In turn, participants, including researchers, journals, editors, publishers to the society in general, can gain rewards from data publication (Costello, 2009; Eysenbach, 2006; Froese, Lloris, & Opitz, 2004; Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011).

Another evidence of data objects becoming the end of itself is that data and its lifecycle become a "common research front" (Meyer & Schroeder, 2015). Such topics include data preservation (e.g., Altman et al., 2009; Beagrie, Beagrie, & Rowlands, 2009; Berman, 2008; Mauthner & Parry, 2009), data curation (e.g., Harvey, 2010; Karasti, Baker, & Halkola, 2006; Lord, Macdonald, Lyon, & Giaretta, 2004; Orchard et al., 2012; Witt, Carlson, Brandt, & Cragin, 2009), and data sharing (Borgman, 2012; Kaye, Heeney, Hawkins, De Vries, & Boddington, 2009; Piwowar, Day, & Fridsma, 2007; Savage & Vickers, 2009; Tenopir et al., 2011), just to name a few.

Software is in a similar position as datasets in this aspect. For example, there is also a strong trend to view scientific software as the ends, not just the means, of computational science (Claerbout & Karrenbach, 1992; Donoho, 2010). Donoho (2010) even goes further to argue that "an article about computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result" (p. 385). Moreover, software is increasingly being studied as the research object in information science (e.g., Howison & Bullard, 2015; Howison, Deelman, McLennan, Ferreira da Silva, & Herbsleb, 2015; D. S. Katz & Smith, 2015; Li, Greenberg, & Lin, 2016; Niemeyer, Smith, & Katz, 2016; Pan, Yan, Wang, & Hua, 2015; A. M. Smith, Katz, & Niemeyer, 2016), which is a broad area to which this proposed work is trying to contribute.

1.3.4 Open science

The notion of science is (or should be) open is as old as modern science itself. It first emerged after science escaped from the paradigm of "secrecy in the pursuit of nature's secrets" (p. 3) did not begin until 16–17th centuries, even though both paradigms coexisted for a long time after that (David, 1998). Moreover, openness (or "communism", to use Merton's word) was summarized as one of the pillars of modern scientific norms by Robert Merton (1968).

As a highly polysemic term, open has many meanings by itself, which is still the case in the expression of open science. Fecher & Friesike (2014) identified five major themes under the umbrella of open science, including infrastructure school, public school, measurement school, democratic school, and pragmatic school. Each theme has its central assumptions about the openness of science, most involved social groups, central aim, as well as tools and methods. What is especially relevant to the topic of this literature review is the democratic and infrastructure schools, both of which aim at making the scientific products openly available.

Data objects, especially data and software, are central to this movement, such as the cases of "open data" and "open software". Both concepts claim that these data objects should be a public good whose rights of reuse and redistribution should not be limited (Molloy, 2011; Murray-Rust, 2008; Prlić & Procter, 2012; Vision, 2010), and thus form a strong basis for the development and study of digital scholarship and cyberinfrastructure, including my future studies.

1.3.5 Distributing science

The "combination of large dataset size, geographic distribution of users and resources, and computationally intensive analysis results in complex and stringent performance" (Chervenak, Foster, Kesselman, Salisbury, & Tuecke, 2000, p. 187) together require that data and computational infrastructure should be distributed. Not surprisingly, the distributed and collaborative nature of knowledge production is one aspect of the definition of the data-driven scientific paradigm (Meyer & Schroeder, 2015).

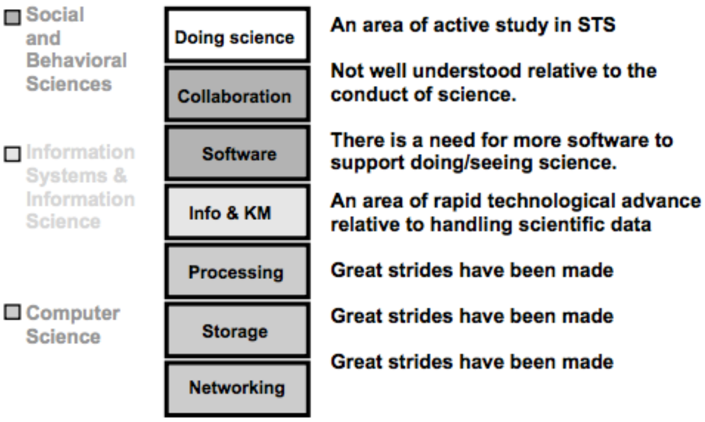
The distributed nature of digital scholarship is reflected in the fact that data objects are more likely to be shared and reused by broader communities. Data sharing has benefits (Kaye et al., 2009; Piwowar et al., 2007), but is not without costs and risks (Bertzky & Stoll-Kleemann, 2009; Borgman, 2012; Foster & Sharp, 2007; Van House, Butler, & Schiff, 1998). Data sharing is also supported by many components of cyberinfrastructure, the most notably data repositories and digital libraries (e.g., Cragin, Palmer, Carlson, & Witt, 2010; He & Nahar, 2016; Tenopir et al., 2011). Moreover, as we will discuss in the next section, many new data-drive research tools are providing more advanced functions for researchers to finish scientific works together, as these tools are becoming distributed in nature.

Scholarship being distributed is clearly a new requirement under cyberinfrastructure. This new attribute of science is changing many aspects of scientific works. The most notable change is that it makes scientific methods more distributed, or what Marres called "redistribution of methods", which suggests a "shared accomplishment" mode of scientific research participated by more actors, such as the case of citizen science and using public data (Marres, 2012). Moreover, as data objects have increased chances to travel between different research communities, even though they can be used by more researchers, they are also more likely to face changed expectations or difficulties of understanding, which affect their reusability (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011). This dilemma is shown in many cases that will be discussed in the next chapter.

1.4 Elements of cyberinfrastructure

This section reviews components of data-centered scientific information infrastructure that are the most relevant to the topic of data and software in scientific activities. The selected components to be reviewed in this section include identification systems, metadata standards, data papers, and the virtual research environments. This is by no means an exhaustive list, which may not be possible at all. But all the components reviewed in this section serve important supportive functions for those tasks that are covered by the next two chapters and my proposed study. What should also be noted is that, scientific data and software, the two key players in this map will be reviewed in the next section.

Before going into each component, one way to overview them is the stacked scheme proposed by Sawyer (2008), where the author identified seven levels of tasks that are supported by cyberinfrastructure, from the most basic layers of networking and storage to the most complicated ones of doing science and collaboration (Figure 3). From this perspective, most of the components reviewed in this section cover the layers above "Info & KM", i.e., they are much more about application than hardware.



Social and Behavioral Sciences
Information Systems & Information Science
Computer Science
Doing science
Collaboration
Software
Info & KM
Processing
Storage
Networking
An area of active study in STS
Not well understood relative to the conduct of science.
There is a need for more software to support doing/seeing science.
An area of rapid technological advance relative to handling scientific data
Great strides have been made
Great strides have been made
Great strides have been made

Figure 1.2: A stacked view of cyberinfrastructure (Sawyer, 2008, p. 358)

1.4.1 Identification system

An important component of any infrastructure is the abilities to identify the elements within this system. Just like what Foucault pointed out: "natural history is nothing more than the nomination of the visible" (Foucault, 2002, p. 144). Cyberinfrastructure is no different from this scenario. The International Standard Book Number and the International Standard Serial Number are just a few identification systems developed before the arrival of cyberinfrastructure to identify publications on the international level. However, as the status of data objects keeps rising, and the limitations of these existing systems to identify data objects are better understood (Lynch, 1998), the development of new identification systems is inevitable.

Digital object identifier (DOI) is the most popular system to identify digital objects so far. It was designed as a persistent and actionable identification and interoperable exchange system, as a response to the needs of managing objects in the cyberinfrastructure (Paskin, 2003). Paskin has identified the following requirements for this system (Paskin, 2010).

- Resolution: the identifier can be used to retrieve the resource per se
- Interoperability: the identifier can be used outside the contexts of the issuing assigner
- Persistence: the identifier denotes the same referent indefinitely
- Uniqueness: one identifier links to one and only one entity

DOI's success is demonstrated by the fact that it has been commonly used for datasets and software entities in various systems, such as DataCite (DataCite International Data Citation Metadata Working Group, 2015), Dryad (Mayo, Vision, & Hull, 2016), and Zenodo (Purcell, 2014), even though this practice is far from consistent, especially for the practice of software citation (Li et al., 2016).

On the other hand, it is just one of the approaches to identify digital objects. There are other systems competing with DOI for digital objects (Duerr et al., 2011), not to say that there are a lot of other types of entities in cyberinfrastructure that need to be identified, such as authors, organizations, and the relationship. During the past few years, there have been various projects aiming at developing the identification system to cover these various types of objects, such as Open Researcher & Contributor ID (ORCID, [Haak, Fenner, Paglione, Pentz, & Ratner, 2012]) and the International Standard Name Identifier (ISNI, [MacEwan, Angjeli, & Gatenby, 2013]) for author names, DataCite metadata scheme to record the relationship between datasets (DataCite International Data Citation Metadata Working Group, 2015; Starr & Gastl, 2011), and various efforts investigating the organization identifier system (Fenner, 2016).

Despite the challenges to name data objects because of their dynamicity (Dourish, 2016; Pröll & Rauber, 2013, 2014), the identification system is arguably the most fundamental component of the whole cyberinfrastructure. Even though the identification system per se will not be the focus of my future study, the abilities to name as many related entities as possible will be the prerequisite for the success of my studies. For this reason, this topic will be enmeshed in the discussions of the rest of this literature review.

1.4.2 Metadata standards

Putting besides "data about data", a stricter definition of metadata could be "structured data about an object that supports functions associated with the designated object" (Greenberg, 2003, p. 1876). Based on this definition, metadata schemes normally bear two broad categories of functionalities: to represent the resources and to facilitate the desired use of these resources (Day, 1999; Green, 2009; Willis, Greenberg, & White, 2012).

There are different types of metadata to perform various functions around these general goals. Lagoze, Lynch, & Daniel Jr (1996) offered one of the most comprehensive lists, including descriptive, administrative, terms and condition, content ratings, provenance, linkage/relationship, and structural. Their list can be mapped to many other different schemes (Caplan, 2003; Gilliland-Swetland, 2000; Greenberg, 2001).

In the scientific infrastructure, metadata plays significant roles to bridge the gaps between information objects and their users. Similar with other standards, metadata is “making things work together over distance and heterogeneous metrics” (Bowker & Star, 2000). This role is reflected in the fact that the library catalog used to be the dominant starting point for scientists to find academic materials before the digital era (Markey, 2007). Despite the fall of library catalog as a type of research tools, metadata persists. For example, it is commonly perceived that good metadata is one of the most important factors to facilitate the sharing of datasets (Edwards et al., 2011; Ingersoll, Seastedt, & Hartman, 1997; Tenopir et al., 2011).

How metadata is designed and implemented is a matter of local contexts. An institution may or may not need a national metadata standard to describe their materials, and their local expertise, workflow or resource may or may not be able to support the adoption of such a standard (Tenopir et al., 2011; Vogel, 1998). Edwards et al. (2011) called this phenomenon “metadata friction”, which refers to the extra efforts that will be taken when a metadata scheme moves beyond a specific community. This gap between metadata scheme and their target users can be seen in those cases where scientists rely upon various resources other than metadata to find and verify a second-hand dataset (Faniel & Jacobsen, 2010; Van House et al., 1998; Zimmerman, 2008).

In terms of this proposed study, metadata plays important supportive roles that are similar with identification systems: it enables us to manipulate and analyze data objects in large scales and automatic ways. On the other hand, the proposed study also aims at enriching the metadata world by introducing new metadata standards about the relationship between data, software, and other research objects.

1.4.3 Data paper and data journal

Data paper is an eminent example of the increasing importance of datasets in the scientific enterprise. Chavan & Penev (2011) traced the history and rationale of this new genre of scholarly publications. They defined data papers as a “scholarly publication of a searchable metadata document describing a particular online accessible dataset, or a group of datasets, published in accordance to the standard academic practices” (p. 3). As compared with traditional scientific paper, data papers only aim at describing the dataset rather than reporting a study.

As a “metadata document”, data papers offer a semi-structured description of the metadata elements of the dataset. Chavan and Penev (2011) mapped a common structure of data papers to the GBIF IPT Metadata Profile. Similar efforts can be seen in Chao's study (2015) to map the method information of 24 data papers to three metadata schemes, including the National Environmental Methods Index, the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (CSDGM), and the Ecological Metadata Language (EML).

An important reason why data papers became popular is to incentivize researchers to share their data with rich metadata (Chavan & Penev, 2011; Rushby, 2015). By introducing research datasets into the peer review process, these datasets are more likely to be found by other researchers through regular academic publication channels, and their publishers will arguable be better rewarded for their efforts to collect and cleanse the data and prepare for the metadata.

As data papers become more popular, there emerged journals dedicated to this new academic genre. In their survey conducted in 2015, Candela, Castelli, Manghi, & Tani (2015) identified 116 data journals, even though only seven of them publish data papers only. Most of these journals are in health science, which is supported by the observations about data papers made by Belter (2014).

A parallel genre of scientific papers called software paper is also gaining momentum (Hong, Hole, & Moore, 2013). One can easily spot a lot of similarities between a data paper and a software paper. For example, both genres serve as a metadata document to describe the data objects and aim at solving the issue to bring the data objects into the peer-review processes.

Data papers (and software papers) clearly extend the traditional definition of a scientific paper, which highlight the rising position of data objects. However, there have been a few warnings about this type of academic publication. First, we are yet to understand the balance between the double mode of production discussed by Knorr (1981), which will be discussed in Chapter two. Second, it is pointed out that data objects serve different purposes in scholarly communication from scientific papers; thus, treating data as publication might shadow the chance to explore new models in accordance with the new scientific paradigm (Borgman, 2015, p. 14).

Because of its importance for the topics of this literature review, data papers will be one of the major targets of my future studies. Potential research questions about this type of academic publication will be raised in the next chapters.

1.4.4 Virtual research environment

Another new form of scholarly communication that is derived from the "fourth paradigm" is the virtual research environment (VRE). In the Report of the Working Group on Virtual Research Communities for the OST e-Infrastructure Steering Group, a VRE is defined as a "set of online tools, systems and processes interoperating to facilitate or enhance the research process within and without institutional boundaries" (Borda et al., 2006, p. 3). Candela, Castelli, & Pagano (2013) identified the following features from VREs: 1) a web-based working environment; 2) being tailored by the needs of a community of practice; 3) offers the whole array of commodities to accomplish the goals of the community; 4) being open and flexible; and 5) promotes fine-grained and controlled sharing of all kinds of research results. Because of these features, VREs are a powerful solution to many requirements of cyberinfrastructure discussed above, such as openness and distributedness.

There is a great number of VREs that have been developed for specific projects covering a broad array of knowledge domains that span medical and biological science (Ahmed, Rodie, Jiang, & Sinnott, 2010; Barga, Andrews, & Parastatidis, 2007; Sinnott & Stell, 2011), natural science (Myers, Trevathan, & Atkinson, 2012), social science (T. Myers, Trevathan, & Atkinson, 2012), and humanities (Bowman, Crowther, Kirkham, & Pybus, 2010; Neuroth, Lohmeier, & Smith, 2011; Rains, 2011; Sarwar, Doherty, Watt, & Sinnott, 2013; Steiner et al., 2014). But as VREs also bring new possibilities of collaboration. As a result, there are also some VREs that are for general uses, the most famous of which include myExperiment (De Roure et al., 2008; De Roure & Goble, 2007) and the Jupyter Notebook (Kluyver et al., 2016; Ragan-Kelley et al., 2014), previous known as the IPython Notebook.

VREs offers a space where a variety of data-driven experiments are integrated into a single system, which not only makes it easier to finish all the tasks in one interface, but also makes it possible for such activities as well as all objects, be they final products or interim products, to be traced and reported. Because of its characteristics and creativity, it will also be a central topic of future studies.

2 Data and software in the research lifecycle

This chapter discusses, in more detail, the relationship built around the two key actors, research data and scientific software, in the scientific information infrastructure. Our review begins with a thorough examination of the definitions of these two concepts, especially their epistemological connections to scientific studies. The relationship between data, software, research method and theory is discussed under a framework derived from STS, which suggests that scientific knowledge is produced as scientific objects are transformed into writings.

2.1 Data and its epistemologies

As an extremely common word used in both academic and public discourses, data is both self-explanatory and extremely difficult to be defined. There are a few examples of definitions of data without referencing other types of information objects, one of which is the definition proposed by the Consultative Committee for Space Data Systems (CCSDS) in their Reference Model for an Open Archival Information System:

> "A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen." (CCSDS, 2002, p. 1-9)

Another example was offered by the National Research Council:

> "Data are facts, numbers, letters, and symbols that describe an object, idea, condition, situation, or other factors." (National Research Council, 1999, p. 15)

Both definitions focus on the representational nature of data, with or without other functionalities and examples of data discussed. Besides these definitions, there is another approach to the definition of data, which is through the comparison between data and other similar concepts. Under this category, Marcia Bates supplied a unique perspective by categorizing the following two types of data:

- Data 1: "that portion of the entire information environment available to a sensing organism that is taken in, or processed, by that organism";
- Data 2: "information selected or generated by human beings for social purposes" (Bates, 2005, p. 14-15)

Data 1 is between what she defined as Information 1 and Information 2: it is a subset of all possible forms of information (Information 1), and has the potential to become a collection of information that has meanings given by human beings (Information 2). Data 2, on the other hand, refers to the information that are selected or generated, including the information generated using scientific methods or collected for social purposes.

Bates' views of the relationship between data and information are quite different from those in the model of "data-information-knowledge-wisdom hierarchy", or DIKW. Some earlier discussions of this model can be traced back to the end of the 1980s (Ackoff, 1989; Zeleny, 1987). This model sets data as the bottom layer of the hierarchy. As a type of raw and unprocessed material, it serves to create more abstract information objects, such as information and knowledge. Sharma (2004) demonstrated that this notion of data was expressed in the literary works of T.S. Eliot and the musician Frank Kappa, which suggests how long this idea was rooted in our collective subconsciousness about the information universe.

These different definitions of data strongly echo what Borgman reminded us, that data is always epistemological (Borgman, 2015). The epistemology of data forms a frame through which certain aspects of the information universe are included/excluded from our focus. Besides its intellectual importance, the epistemology of data has important practical implications. For example, Wynholds (2011) discussed four functions of data identity played in scholarly communication:

> "1. datasets be represented as semantically and logically concrete object; 2. the identity of the dataset is embedded, inherent and/or inseparable; 3. the identity embodies a framework of authorship, rights, and limitations; and 4. the identity translates into an actionable mechanism for retrieval and citation." (p. 218)

To summarize, how a data object is perceived has significant impacts on if and how its functions in scholarly communication are viewed and fulfilled. Two epistemological topics about data will be discussed below, namely data is raw and unprocessed and data is the means of scientific studies. We are not trying to adopt any stance concerning these two topics. However, what is important is that various stances exist around these two and other epistemological topics in individual research communities, which determine how data objects are created, curated, shared, as well as how they are connected to scientific studies. It will be one of our tasks in the propose study to identify how a community perceives the epistemological relationship between data and research, and how these perceptions affect the lifecycles of these data objects in the community.

2.1.1 Data as raw and unprocessed materials

The view that data is raw and unprocessed, as expressed in many definitions reviewed above as well as common expressions like "raw data" and "primary data", has been a basic way in which data has been perceived from ancient times. Being categorized as a "foundational" view by Hammarberg (1981), it is inscribed in the Latin etiology of the term "datum", which is "that is given prior to argument" (Rosenberg, 2013, p. 36).

Despite its popularity, this stance has been increasingly challenged by an alternative view in modern scholarship, that data is theory-laden, which means that data does not exist before or without scientific theories. This alternative view has been promoted by researchers from multiple traditions. For example, Norton & Suppe (2001) proposed the concept of data-model symbiosis, which suggests that data relies upon modeling that is built into instrumentation. Building upon the discussions of Norton and Suppe, Edwards

(2010) argued that neither data nor model is pure: they helped to form each other from the very beginning of either lifecycle. Moreover, data also needs theories to be evaluated so that it can enter any information system (Da Costa & French, 2003; Hammarberg, 1981)

Moreover, the foundational view of data is also highly problematic in fields where interpreticism and the exposure of situatedness of researchers have higher priority (Schöch, 2013). Johanna Drucker proposed to use the word "capta" to replace "data", which means that data is taken and that data-taking happens in contexts (Drucker, 2011). More recently, Denis & Goëta (2014, 2017) offered a case study of open government data, which shows that the rawness of data is not only an increasingly important requirement in the community, but also a prerequisite for data to be open; in many cases, government datasets must to "re-rawified" so that they can be shared.

The changed levels of rawness are an important aspect of the lifecycles of data objects. What we can learn from the stories mentioned above is that as datasets go through their lifecycles, their rawness is not always decreasing, and is not necessarily aligned with other lifecycles, such as those of scientific studies. The level of rawness could have significant implications for if and how the data can be used, which makes it an important variable to be pursued in our future studies. Moreover, when conducting such studies, it is important to take a phenomenological standpoint, by using the strategies such as infrastructural inversion (Bowker & Star, 2000) and data journey (Leonelli, 2014) to track the lifecycles of data objects per se.

2.1.2 Data as means of scientific studies

What is also implied in the Latin origins of "datum" is that data is a factual object that leads to knowledge. After his historical examination, Rosenberg (2013) concluded that "facts are ontological, evidence is epistemological, data is rhetorical" (p. 18). One of the corollaries of this statement is that data is means rather than ends of scientific studies.

This view is unmistakably shown in the research lifecycle. In a typical research lifecycle (e.g., Vaughan et al., 2013), research starts from scientific ideas; through the collection and analysis of data, the results are produced and reported. Of course, this highly rudimentary model reflects a deductive approach of science. As discussed in the previous chapter, big data, at least at its face value, represents a rise of inductivism (Erevelles et al., 2016; Frické, 2015; Krumholz, 2014), which is clearly a challenge to this epistemological stance, even though the inductivism-deductivism relationship in big data research is still a highly disputable question (Kitchin, 2014; Leonelli, 2012).

The bigger challenge to this stance is that as data moves to the center of the scientific enterprise, it is becoming the ends of itself, as is discussed in the previous chapter. But this perception, as other epistemological stances discussed above, is hugely variant across different research communities, not to mention knowledge domains (Borgman, 2010).

2.2 Scientific software

Scientific software is a central component of cyberinfrastructure and computational science. As compared with data, the difficulties to define software do not lie on its commonality and ambiguity, but the multiplicity of terms that are used to express similar meanings. This section will review the definitions of a set of software concepts, as well as the typologies of scientific software.

2.2.1 Definition of software concepts

Different terms are used to describe the complex of concepts around the phenomenon we call software. This section reviews the concepts of software, code, and algorithm.

On the top of the pyramid of this conceptual complex is the term software. In an authorized definition offered by the IEEE Standards Coordinating Committee (1990), software is defined as computer programs and procedures as well as any associated documentation and data pertaining to the operation of a computer system, as contrasted to hardware and firmware. As we will see in the discussion of other concepts, this concept is broad enough to cover all related situations of software.

Behind this general concept, code denotes to the "textual form of programming code" (Berry, 2016, p. 29) that forms the functional backbone of any software. There are two major types of code formats based on which software is built: an executable code (object code), or the code after compilation, and source code (Kennedy, 2001).

Different from code, an algorithm forms a more abstract layer of instruction of "computational procedure" (Cormen, Leiserson, Rivest, & Stein, 2009, p. 4) implemented by the code, by specifying the tasks to be accomplished as well as the outputs and inputs of each step. It can take more forms, such as a formula, a set of rules or steps, or simply an approach to solving a problem (Knorr-Cetina, 2016). Thus, algorithms are not necessarily software (Gillespie, 2014); algorithms are both more than programs (algorithms are free from the "material constrains") and less than programs (in the sense that programs contain non-algorithm material) (Dourish, 2016, p. 2).

Despite the variance of their semantic meanings and positions in the conceptual complex of software, all these concepts will be largely used as a group during the rest of this literature review, with their individual differences ignored. They, collectively, serve as the engine of scientific tasks and the contemporary ecosystem of information. Software entities are designed to process data, but their relationship with data are variant depending on the types of data and software, and the contexts of the use of data and software.

2.2.2 Scientific software and its classification

An obvious way to define scientific software is to specify what attributes science adds to software, such as:

> “[S]oftware with a large computational component and provides data for decision support” (Kelly & Sanders, 2008, p. 1)

> “[A]pplication software that includes a large component of knowledge from the scientific application domain and is used to increase the knowledge of science for the purpose of solving real-world problems” (Kelly, 2015, p. 50).

Both definitions depict scientific software as application software with certain attributes to support scientific work, either because of its computational capacities or its knowledge elements.

However, the discussion of Kelly, Smith, & Meng (2011) is closer to how scientific software will be examined in our contexts, where the authors define two types of scientific software, namely end-user application software and “tools that provide support for scientists to express their models in code and execute their software solutions” (p. 7). These two scenarios are exemplified by the distinction between built-up software and software that supports user-created source code.

The other way to define scientific software is through its functional classification. However, because of the complex nature of scientific tasks, such classification only exists in highly localized contexts such as software repositories. GAMS (Guide to Available Mathematical Software) Classification Scheme is one of the most popular scientific software classifications, which is specifically designed for mathematical software (Boisvert, Howe, & Kahaner, 1983). It defines 20 categories of mathematical problems. Under each category, subclasses are applied. Some other efforts in the same category include those of di Serafino, Maddalena, Messina, & Murli (1998) and Rice (2013). More broadly, a classification system named Taxonomy of Digital Research Activities in the Humanities (TaDiRAH) was developed for general research tasks (as well as objects) centered on digital humanities and has been implemented in a few digital projects, such as the Digital Research Tools (DiRT) repository (Borek, Dombrowski, Perkins, & Schöch, 2016; Perkins, Dombrowski, Borek, & Schöch, 2014). In our future studies, an important task would be to apply such a classification scheme to the analyses of data-software relationship, be the scheme a domain view or a task view.

2.3 Laboratory studies

An especially important school of thought that shaped contemporary studies of science, including laboratory studies, is actor-network theory (ANT). ANT focuses on how technical artifacts are constructed from the heterogeneous networks formed by human and nonhuman actors (Law, 1992). Under its influences, laboratory studies, initiated by the trailblazing work of Latour and Woolgar (1979), take a strong focus on how scientific knowledge is produced by observing the real scientific activities. As noted by Woolgar in his review (1982), a central feature of this type of study is to offer a description of science “as it happens” (p. 483). Science is taken as artifacts are constructed through the constant interactions between various actors, rather than facts. This approach, based on Latour’s explanation, is opposite to the traditional scientific epistemology to treat science as black boxes:

> “The word black box is used by cyberneticians whenever a piece of machinery or a set of commands is too complex. In its place they draw a little box about which they need to know nothing but its input and output… That is, no matter how controversial their history, how complex their inner workings, how large the commercial or academic networks that hold them in place, only their input and output count.” (Latour, 1987, p. 2)

What is the prerequisite for the pursuit of such complex interactions, based on the agenda set up by Latour and his colleagues, is the translation of standpoints between different actors (Law, 1992). Callon, Courtial, Turner, & Bauin (1983) defined translation as “all the mechanisms and strategies through which an actor… identifies other actors or elements and places them in relation to one another” (p. 193). In his book *Science in Action*, Latour further discusses some strategies of translation, which include catering to other people’s explicit interests, persuading others to take a detour, reshuffling interests and goals, and becoming indispensable (Latour, 1987, p. 108–121). As in the case of the history of scientific works around TRF(H) described by Latour & Woolgar (1979), TRF had different meanings by various actors in the beginning; a scientific consensus is only possible through a series of negotiation and persuasion among these players, before some voices prevail others, which is when some scientific statements are accepted as scientific facts.

An important reason why ANT is selected as a foundational theory for this literature review, besides its historical connections to studies on laboratory activities, is that ANT puts non-human objects in a more important position in the construction of science. This approach could shed stronger light on the materiality of objects and how the materiality could affect how science is conducted.

In the tradition of laboratory studies, the laboratory space and scientific publication are two ends of scientific studies. This view is rooted in Latour and Woolgar (1979); as an ethnography of the “laboratory life”, the authors identified the laboratory space as a “literary inscription” apparatus, which “transform pieces of matter into written documents” (p. 51). Based on this notion, the system of experiment and that of scientific writing will be discussed below, as the basis of the rest of discussions in this chapter.

2.3.1 Experimental system

Experiments are an integral part of laboratory studies. Following Knorr Cetina's definition (1999), experiments “conduct ‘science,’ while laboratories provide the (infra-)‘structure’ for carrying it out” (p. 42); as a result, experiments are the core areas where causal events are attached with signs that are either visual or auditory, during which process scientific objects serve as the representations of the nature to produce scientific knowledge.

In his book *Toward a History of Epistemic Things*, Hans-Jörg Rheinberger (1997) focused on the material side of the laboratory space: how material objects are processed in the experimental system. This topic represents a point that is one step further from the tradition built by Latour, Woolgar and others; that is, "[genesis and development of scientific facts] amounts to a relation between objects themselves" (Rheinberger, 2005).

Rheinberger identified two types of objects in the experimental system. The first is called "epistemic thing" or scientific object, which is the research objects that are the target of scientific inquiries. The second is called technical object, which is the experimental conditions through which research objects "get entrenched and articulate themselves in a wider field of epistemic practices and materials cultures, including instruments, inscription devices, model organism, and the floating theorems or boundary concepts attached to them" (p. 29). The experimental system can only be operated with the interactions between these two types of objects. Moreover, Rheinberger argued that their relationship is highly entangled. First, scientific objects are embedded in and subject to the conditions in which they are studied. Second, both types of elements participate in a "non-trivial interplay, intercalation, and interconversation, both in time and in space" (p. 29). Last, just like the relationship between infrastructure and work, there is no clear boundary between these two types of objects: their difference is functional rather than structural. Or to borrow the author's own words, "technical objects are… the frozen product of former epistemic activity" (Rheinberger, 2005).

Thanks to his strong focus on the materiality of scientific practices, this complex relationship between both types of objects depicted by Rheinberger offer a nice framework to study roles played by different types of scientific data objects, and their potential interactions. The relationship between data and software is arguable similar with that between scientific objects and technical objects, both in terms of their different functions in scientific studies and their blurry boundaries. As such, it would be interesting to pursue Rheinberger's research agenda focusing on these data objects, especially:

- How scientific knowledge flows from scientific studies to software?
- How scientific knowledge "frozen" in scientific software is reused and updated in their later uses?
- How Rheinberger's theory can be applied to a knowledge domain other than physical science?

For the last question, Freud offered one example of extending such STS models developed from "hardcore" science into other knowledge domains; he argued that psychoanalysis has a similar mode of scientific operation, in the sense that a space, independent from the real world, is created where research objects are processed, even though the nature of scientific objects in psychoanalysis is very different from other areas (Knorr-Cetina, 1992). By definition, the experimental system should not be limited to just natural science or even computational science. Interdisciplinary pursuits will be continued and extended in my future studies to prove this assumption.

2.3.2 Norms of scientific writing and scholarly communication

Scientific publication is the other end of the scientific action defined by laboratory studies. Latour and Woolgar (1979) categorized the function of the laboratory space to be producing scientific texts from the objects. This function makes the laboratory an “inscription device” aiming at persuading readers by the documents created from scientific studies. These documents, argued by Latour (1990), are “immutable mobiles”, because they are easily transportable across space and time (like through publication or sharing with other people) with (relatively) fixed meanings.

This production-driven nature of scientific objects has been observed by many researchers. One of such examples is Larraine Daston (2000), who stated that scientific objects are never inert but always lead to the production of scientific knowledge, by “producing results, implications, surprises, connections, manipulations, explanations, applications” (p. 10).

What happens in the experimental system needs to be translated into texts, in order to accommodate different characteristics of both spaces. On the one hand, scientific studies in the reality are full of uncertainties; scientists need to make contingent decisions at every step of the experiment, which is also affected by various extra-scientific factors. On the other hand, one of the highest requirements for scientific writing is the certainty that can display creativity and defend criticism. Knorr (1981) argued that scientific papers can only be finished with a double model of production: the instrumental mode, which is to decontextualize the results created from the laboratory activities, and the literary mode, which is to recontextualize the results based on the new territory and niche; what connects these two modes is rule of transformation:

> “Avoidance of reason and the typification of the paper’s version of Method converts the painfully constructed ‘way’ (or method) of the laboratory into a natural consequence of the work’s overall purpose and the reasoning contained in the introduction.” (p. 118)

After such transformations, technical details are largely pruned off in the final scientific writings, and they are interwoven into reasons neatly so that to persuade the readers (Swales, 1990). These reasons, in many cases, did not exist when the experiments were conducted. As a matter of fact, such requirement of scientific writing is not only personal strategies, but also community norms in many cases (M. J. Katz, 2009). Another explanation of the removal of technical details is that every reader of the scientific paper is assumed to be able to understand the contexts of the experiments, or like what Borgman (2015) argues:

> “Details necessary to replicate the study are often omitted because the audience is assumed to be familiar with the methods of the field.” (p. xviii)

However, an undisputable consequence of this series of transformation is that memories of scientific processes are lost: scientific papers become “residual descriptions” (Knorr, 1981, p. 130) of the highly contingent and recursive laboratory activities. As the “story of an ideal past in which all the protocols were duly followed” (Bowker, 2007, p. 25), it is extremely challenging, if not impossible at all, to recreate the experimental processes from the textual records. However, to reproduce scientific results is becoming a more urgent requirement in data-driven scholarship, based on our discussions in the previous chapter.

One of the key questions inspired by laboratory studies, that we will pursue in the future studies, is how the requirement of reproducibility is differently fulfilled by classic scientific papers and the new genre of data papers. In traditional scientific papers, technical details are often ignored, at least in the traditions of natural science. However, we would like to extend such studies to other knowledge domains that are also computational. But more importantly, we would also like to examine if and how such a process of knowledge production is represented differently in data papers compared with scientific papers. It is assumed that in data papers, more details about the creation and cleaning of data objects should be offered. Yet, how such details are compromised by the writing norms of this genre is a rarely examined research question so far.

2.4 The complex of data-software-method-theory

The relationship between research data and scientific software in scientific activities is almost self-evident, because after all, "data doesn't do anything of itself" (Berry, 2011, p. 51). Both entities make their unique contributions during the process of scientific knowledge production. However, as much as we cannot understand data without software, or vice versa, we also cannot fully understand any of them and their relationship without considering other entities in the scientific knowledge infrastructure. For example, scientific software is so strongly connected to and in many cases defined by research methods and scientific theories. Following this path, this section will pursue the relationship between data and software, by juxtaposing them with research methods and scientific theories. One way to define these relationships is through the concept of package developed from STS, which will be discussed in the beginning of this section. We will then address various possible packages composed from these entities, and specify how these packages

2.4.1 Package of scientific works and the theory-method package

One way to understand the relationship between data and software is through the concept of package developed from the tradition of STS. Fujimura (1987) defined a package as how tasks are organized into standardized procedures, with the goal to increase the doability of the tasks; in the scientific practice, a task is doable when it meets the requirements on the levels of experiment, laboratory and social world at the same time. STS researchers have argued that scientific works are procedures saturated with uncertainties (Fleck, 1981; Kuhn, 1962; Star, 1985). Based on this view, Fujimura (1987) claimed that packages make use of the methods of modularization and standardization to reduce the uncertainties in scientific works; this is achieved by cutting a question into smaller units, which can be integrated into standardized procedures so that they are easier to be solved than the bigger problem.

Even though many scientific entities can be packaged together, the packages that are comprised of theories and methods ("theory-methods package") are the most frequently discussed in STS studies. This type of package is composed of a "scientific theory and a standardized set of technologies which is adopted by many members of multiple social worlds" (Fujimura, 1992, p. 169). In order for the package to succeed, its ontology, epistemology, and practice should be integral and co-constitutive. For example, Star described the historical processes of how the theory about brains segments and nervous system were successfully localized in England and broadly accepted as a fact in the early 1900s; she recorded how scientists from different lines of research co-developed theories, methods, and laboratory procedures around this topic through a collective process of action, and gradually reached the final breakthrough, and how the success is not possible without strong organization supports which makes social factors inevitable in these theory-methods packages (Star, 1989).

The concept of package is significant for understanding scientific activities. It is a relevant concept to cyberinfrastructure because cyberinfrastructure should not change the nature of scientific works thoroughly as it replaces previous scientific information infrastructures. Moreover, the applicability of the concept of package to the topic of this literature is reflected in the fact that package is also a popular metaphor in information and communication technologies (ICT), such as the metaphor of "computer as package":

> "[T]he package metaphor describes a technology that is something more than the physical device. In the case of computing, the package includes not only hardware and software facilities, but also a diverse set of skills, organizational units to supply and maintain computer-based services and data, and sets of beliefs about what computing is good for and how it may be used efficaciously." (Kling, 1980, p. 79)

Similarly, "software package" is also frequently used in all kinds of narratives. It normally contains much more than just codes, as is the case of how R packages are defined (Wickham, 2015).

From a methodological perspective, most studies using this concept are based on qualitative methods; few of them tried to translate the concept of package into quantitative and/or scientometric terms, which is a major methodological motivation of this proposed study. Previous efforts have been done to translate concepts between STS and scientometrics, such as mapping the concept of invisible college to scientometrics terms (e.g., Gmür, 2003; Lievrouw, 1989; Noma, 1984). Thanks to the growing text techniques accompanied with citation analysis, we are in a much better position to deal with this issue. One assumption that we can make about such packages is that entities belonging to them are supposed to show up in closer positions with specific textual patterns around their locations. Thus, we might be able to track these patterns as references to these packages through using co-mention networks of different scientific entities and NLP techniques. The technical background of this method will be discussed in more detail in the next chapter.

The rest of this section reviews some potential packages and/or important relationship between data, software, scientific method, and theory.

2.4.2 Data and software

The first set of relationship exists between data and software. Data and software are two distinct classes of objects in scientific activities; each plays different roles and displays different characteristics. For example, Katz et al. (2016) summarized the differences between software and data in terms of scholarly communication:

- Software is executable, data is not
- Data provides evidence, software provides a tool
- Software is a creative work, scientific data are facts or observations
- Software suffers from a different type of bit rot than data
- The lifetime of software is generally not as long as that of data (p. 2-3)

Some points mentioned above reflect epistemological stances previously discussed in this chapter. What is true in the discussions of Katz and his colleagues is that data and software, as two categories of data objects, enjoy different lifecycles and perform various roles in scientific studies. On the other hand, this conclusion should not shadow the fact that each instance of datasets and software are unique and form distinct relationship with each other.

Moreover, we should also remember that software is technically data; we can benefit from analyzing code as data by using techniques and theories from information science (Marcus & Menzies, 2010). Moreover, it would be of great help if software is curated and preserved like datasets (Lynch, 2014). To make this issue more complicated, research data and scientific software enjoy a highly entangled relationship in the cyberinfrastructure. Gillespie (2014) correctly pointed out that algorithms (or any other software entities) must be paired with data to function; these two entities are different but economically and ideologically concert. This relationship is well represented in the title of the book written by Niklaus Wirth, *Algorithm + Data Structure = Programs* (Wirth, 1978). This statement illustrates not only the co-constructive relationship between data and software, but also the "ontology of the world according to a computer" (Manovich, 1999, p. 84).

Most importantly in the context of this section, based on the concept of package, it is undeniable that datasets and software are constantly packaged together in scientific activities. For example, a piece of software is likely to be physically packaged with one to many datasets, which serve as exemplified data for the functions performed by the software. It is also common that certain types of packages and software are more likely to be used together or described in the literature. A few questions about this type of scientific packages formed between datasets and software entities are yet to be answered, such as:

- How are these various data-software packages formed? What are the factors that contribute to their packging, especially the mediation of other scientific and extra-scientific objects?

- How are these packages be used or reused in scientific activities, as a collective entity or individual entities? Is there any different pattern of usage determined by the people who use them?
- How are these packages represented in scientific texts?

2.4.3 Data objects and theoretical objects

Based on the dichotomy between inductive and deductive reasoning, theories are either the beginning or the end of scientific studies, which interact with data objects in different ways in research lifecycles. In most cases, such different interactions take the form of what specific software and data entities are used, and various details about how they are used. But from the perspective of the data lifecycle, theories also affect multiple steps of how data is collected, manipulated, shared, and reused.

However, scientific theory itself is not a monotonous concept. Especially in physical science, we are constantly facing the subtle differences between scientific theories and models, both of which serve to represent the empirical world. In its classic meaning, scientific models are the bridge between scientific theories and the physical world. As Hacking (1983) observed, there is often an untranslatable relationship between a theory and the experimental data; as such, models serve to translate data or observations into theories in the logical space. Based on Suppes' theoretical framework, Fraassen (2008) identified three types of models for data to be "abstracted into a mathematically idealized form" (p. 167), namely the data model (which summarizes the relative frequencies), the surface model (which idealizes the frequencies into continuous values), and the theoretical model (by fitting the idealized values into a theory).

It is worth noting that most of the discussions mentioned above are derived from natural science. It would be one of the goals of the proposed study to examine their applicability to other knowledge domains. Moreover, it is also within the broad goals of my future studies to examine the relationship between data objects and theoretical objects in both the actual scientific processes and scientific writings. From a scientometric perspective, following the general strategy we have discussed in section 2.3.2, theoretical objects will be treated as a parallel type of objects to be identified in scientific publications. But a bigger task would be to identify their relationship with other data and software objects and the interactions between data and software. Such questions include:

- What are the relationship between theoretical objects and data objects in both research lifecycles and data lifecycles in both laboratories and scientific writing?
- What are the relationship between theoretical objects and the relationship between data and software entities, if at all?
- How are these relationships described differently between scientific papers and data papers?

Such investigations will be conducted in the context of scientific activities in the laboratory space and scientific representation in the writing space. Each space is expected to give us specific insights about these research objects and their relationship. Moreover, the comparison between studies conducted in individual space will help us better understand how knowledge production is the same or different in the data-driven paradigm as compared to more traditional scientific paradigms.

3 Citation theories and their applications on data objects

This section traces the developments of the theory of citation and evaluate the potentials of applying this broad family of theories and related methods to digital data objects. In the context of this literature review, we believe the essence of citation is the inter-document relationship, or, in the context of data-driven scholarship, document-object relationship. In this sense, we are not following the citation/reference distinction in the literature (Egghe & Rousseau, 1990; Narin, 1976; D. J. Price, 1970); rather, these two terms are used interchangeably in this document.

Even though there does not seem to be a consensus concerning the exact origin of citing other's works in scientific writings, researchers agreed that reference has a long history and has been closely connected to the modern scientific enterprise from its dawn until today (De Bellis, 2009; Grafton, 1997; Neville, 2010; D. de S. Price, 1963).

In this chapter, we examine a few families of citation theories, and what we can learn about digital objects from these theories and the methods that are derived from them. Theories are important to citation studies given the fact that there have been multiple calls that we are lack theories to interpret citation data (Cronin, 1981; Edge, 1979; Gilbert, 1977; Kaplan, 1965; Luukkonen, 1997; Zuckerman, 1987). As will be discussed later, these theories help to build methods to conduct studies, which in turn create results that may support or resist these theories.

The first part of this section focuses on the three major citation theories that have been established, namely the normative theory, the rhetoric theory, and the linguistic theory. For each paradigm, we examine how it was influenced by the broader trends in social science, how it influenced more specific research methods and the interpretation of empirical evidence, and at the same time how it was supported or resisted by these evidences. Moreover, an important method, content and context analysis is discussed, not only because it is connected to all the major theories we review earlier, but also because it will be an important method in the proposed study.

The second part of this section examines how the data-driven research paradigm reviewed in previous sections is connected to this area. More specifically, we talk about how citation studies have helped us understand the nature of data objects and their relationship, and how these studies are connected to the traditional citation analysis theories and methods as well as the new information infrastructure.

3.1 Three "paradigms" of citation theory

The shift of the theories of citation in the history of this field is similar with what Thomas Kuhn called "paradigms" (1962). Even though the concept of paradigm was not precisely defined by Kuhn and used by him in multiple ways (Ingram, 1993; Masterman, 1970), it was broadly defined as to denote accepted models or patterns that involve conceptual, theoretical, instrumental, and methodological commitments that needs continuous articulation. The replacement of paradigms forms the basic structure of scientific revolution (Kuhn, 1962). However, even though the different ideas about citation reviewed in this section are distinct models, they are hardly as comprehensive as Kuhn's scheme: they complement as much as replace each other. We will use the term paradigm in this section in a more general sense than that was used by Kuhn.

3.1.1 The normative theory

As one of the founding fathers of both sociology of science and citation analysis, Robert Merton's normative theory of science is one theory that holds a unique position on the boundary between sociology of science and information science. This theory is built upon the notion that individual scientists are dependent on and inevitably deeply influenced by the social structure of science in significant ways.

His most famous contribution to scholarly communication is an essay titled "Science and Technology in a Democratic Order" (Merton, 1942). This paper was first published in Journal of Legan and Political Sociology in 1942, and was reprinted in his later books with different titles (Merton, 1968, 1973). In this essay, Merton discussed the importance of studying the values and norms of modern science, because these values and norms are some of the most important components of how science is defined, besides the scientific methodology and the accumulation of scientific knowledge. Moreover, these values are institutional legitimatized, acquired and internalized by scientists, and formed one's scientific conscience.

Merton identified four elements of these institutional imperatives, including universalism (truth should be examined by impersonal criteria), communism (scientific findings are owned by the community of scientists), disinterestedness (institutional control of the motives towards the common benefits of the scientific community), and organized skepticism (scientific claims are subject to scrutiny before being accepted).

His normative theory had strong influences on the development of citation theories. One of the earliest uses of the normative theory in citation studies was offered by Kaplan (1965), just one year after the launching of Scientific Citation Index (Garfield & others, 1964). Kaplan stated that the major function of citation is the "reaffirmation of the underlying general norms of scientific behavior" (p. 181). Moreover, Kaplan introduced the metaphor of property to describe the nature of scholarly communication and the roles citations play in this system: citing is to pay intellectual debts. This idea is closely connected to Merton's four norms, but especially the norm of communism.

Kaplan's use of Merton's theory was accredited by Merton himself in his foreword to Eugene Garfield's book *Citation Indexing* (Garfield & Merton, 1979). In this article, Merton affirmed the metaphor of property. To go even further, he asserted that the "composite communications-intellectual-property-and-reward system" (p. vi) of scholarship establishes a moral-cognitive framework for citation behaviors. From the cognitive perspective, scientists need to express the historical lineage of knowledge through citations. From the moral perspective, citations are a means to establish the reward system, so that the intellectual debt is repaid. And thus, citation is an important vehicle for the accumulative reproduction of scientific knowledge.

This “citation-as-reward-system” metaphor has been continuously supported by the fact that there is a positive correlation between the quality or scientific significance of a research paper (or entities on other levels, such as author, institution, and journal) and the number of citations it receives. This idea forms the basis of using citation data to quantitatively evaluate the research performance, which has been one of the most important use cases of citation data until today but has also caused great controversies (Borgman & Furner, 2002; Leeuwen, 2005; Narin, 1976).

The assumptions behind the normative theory of citation has been summarized and challenged in many ways. A notable example is the following list offered by Borgman and Furner (2002):

1. that the motivation or goal of the citer is to identify all and only citation-worthy works–works that “ought” to be cited in the citing work;
2. that the general result of citers’ activities is such that (a) all works that ought to be cited in the citing work indeed are cited, and (b) all works that are cited indeed ought to be cited in the citing work; and
3. that the quality of a given citable work consists in its citation worthiness, and thus may be measured by citation counts (p. 12).

The challenges to the normative theory happened early in the history of citation studies. Besides the criticism from more interpretive perspectives, warnings were even raised by researchers who adopted this theory. Kaplan offered the observation that not all citations are rationally given by the authors, neither are all of them reinforcing communism (Kaplan, 1965). This type of criticism was later extended to the point that the normative theory is often lack of considerations of the context, content, and motivation of citation, which greatly inspired studies in the area of content and context analysis. This type of reflexivity is also revealed in Merton’s own study about obliteration by incorporation (Merton, 1988) and Eugene Garfield’s reflections on the superficial interpretations of citation patterns (Garfield, 1979, 1988).

Another major limitation of Merton’s normative theory is its “document-centric view” of scholarly communication, that is, how research products are cited in formal published papers is the central, if not the only, focus of this theory (Small, 2004, p. 75). This view embeds the assumption that citation of a document equals to the use of the document (L. C. Smith, 1981), which ignores the importance of informal scientific communication in the overall landscape of science (Edge, 1979). In a way, this is one important reason why the normative theory was abandoned by the next generation of sociologists after Merton.

Despite all these limitations, the normative theory established the methodological and epistemological foundations of citation analysis. As the dominant paradigm in the sociology of science before the 1980s (Star, 1995), it influenced a broad array of studies in STS and information science. In terms of citation analysis, Merton deeply influenced early researchers in this field such as Eugene Garfield and Henry Small (Garfield, 2004; H. Small, 2004). Science Citation Index, the most important piece of infrastructure of citation analysis, has subtle but strong connections to Merton’s theory (Merton, 1977). Thanks to these passages of influence, the normative theory inspired, directly or indirectly, some important research methods, most notably the content and context analysis, which will be reviewed later this chapter.

The normative theory also received the supports from many empirical studies, especially the correlation between the citation count and scientific significance or quality of a work, (e.g., Clark, 1957; Cole & Cole, 1971; Lawani & Bayer, 1983; Myers, 1970; Virgo, 1977) and the evaluative use of citation data in the form of index (e.g., Ball & Tunger, 2006; Costas & Bordons, 2007; Van Raan, 2006). These studies formed the tradition of this field, and having been having constant conversations with new theories, methods and evidences since then.

3.1.2 The rhetoric theory

As a paradigm in both STS and information science, Mertonian sociology was increasingly challenged during the 1970s, and was eventually replaced by the next wave of theories of science by the beginning of the 1980s. This shift was characterized as a shift from normative or structuralist theories to interpretive theories of sociology (Borgman & Furner, 2002; Law, 1974; Law & French, 1974). One important member of this wave of theories is the actor-network theory (ANT), which has been reviewed in the previous chapter.

Bruno Latour’s studies on laboratory practices also had substantial influences on the construction of citation theories (even though Terttu Luukkonen [1997] argued that Latour’s theory failed to have the same influence compared to the normative theory). In their book *Laboratory Life* (Latour & Woolgar, 1979), the authors responded to the normative theory of citation. By citing earlier criticism on the normative theory (e.g., Mitroff, 1976; Mulkey, 1976), they concluded that the existence of norms in citations is not well-supported by empirical evidences; rather, many studies suggest that scientists are appealed to conternorms or just the eagerness to give a good impression. Moreover, the authors also criticized studies on science “with such macroconcerns” such as “studies of the size and general form of overall scientific growth, the economics of its funding, the politics of its support and influence, and the distribution of scientific research throughout the world” (p. 17). Latour and Woolgar’s objections to both the normative theory and the macro-level analysis based on this theory are based on a similar consideration of the rest of theory book, which is the challenge to the scientific self-evidence.

In terms of scientific citations, Latour and Woolgar focused on the persuasive or rhetoric functions played by citation in scientific communication. These functions made citations acting like an “inscription device, i.e., they are a resource used by researchers to support and defend their knowledge claims. For example, in the case of TRF(H) that we have discussed earlier, researchers use other people’s papers in new contexts, i.e., their own papers, to strengthen their own arguments; these uses are often located in new contexts that are totally different from those where the original texts were written (Latour & Woolgar, 1979).

Latour’s idea about citation is not without predecessors. Nigel Gilbert first proposed the notion that citations act like a rhetorical device. Following this interpretive path, he focused on the roles of citations in the actual processes of research, and noted that "one function of reference is therefore to act as a device which establishes the authority on which the author's argument is founded" (Gilbert, 1976, p. 287). In another paper of his (Gilbert, 1977), he specifically addressed how the new metaphor of citations could better interpret the "various ways in which cited material may be used" (p. 115) which cannot be differentiated by the normative theory. By citing evidences offered by content and context analyses, he noted that authors could cite papers that are important and correct, erroneous, or respected and less relevant, all of which could increase the persuasiveness of a paper in different ways. This phenomenon is also discussed by Latour as a piece of evidence to support his overall arguments:

> "[Sources] may be cited without being read, that is perfunctorily; or to support a claim which is exactly the opposite of what its author intended; or for technical details so minute that they escaped their author's attention; or because of intentions attributed to the authors but not explicitly stated in the text." (Latour, 1987, p. 40)

Even though some of these abnormal patterns of citation have been spotted by researchers subscribing to the normative theory, they are largely ignored in their construction of the citation theory. This, according to Luukkonen (1997), is one of the biggest contributions Latour made to this field.

A more direct contribution in methodology made by actor-network theory is the method called co-word analysis. First proposed by Callon and his colleagues (1983), this method aims at tracing the coappearance of words, as the indication of the "problematic network", to understand how authors identify actors and the interests and strategies of these author, how they define the problem and correspondingly objectify, remodel, and transfer knowledge (p. 193). This concept is consistent with the core idea of actor-network theory treating science as a network of actors constantly define problems and build connection. Moreover, it is also a method staged in scientific texts. Based on the concept of inscription device (Latour & Woolgar, 1979), it takes words as a vehicle of interests and targets how "knowledge is produced by making use of existing texts and acting upon them" (p. 198). This is not a method that directly deals with citations. However, this method is one that parallels to co-citation analysis (Callon et al., 1983), and has inspired later works analyzing the contexts of citation (Leydesdorff, 1998).

3.1.3 The symbolic theory

The last major paradigm of citation is to see citations as signs. This view was first proposed in the paper entitled "Cited Documents as Concept Symbols" by Henry Small (1978). In the beginning of this paper, he challenged some previous citation studies:

> "[T]hey have missed an important and perhaps crucial point. Very little, if any, attention is given in these studies to the scientific content of the citation context… Hence these studies have missed the role citations play as symbols of concepts or methods." (Small, 1978, p. 327–328)

Empirically, the symbolic theory is rooted on many pieces of empirical evidence that inspired the rhetoric theory. Theoretically, Small's idea is influenced by the symbolic theory developed by Edmund Leach, a British social anthropologist, who is deeply connected to Claude Levi-Strauss. Based on Leach's definition, this symbolic view means that an object stands for an idea; thus anthropologists should pursue the connections between phenomena, symbols and meanings, as is the case of his most famous book concerning the kinship in Burma (Leach, 1954). In the case of citations, Small (1978) argued that the cited document is the object standing for the idea that is expressed in the citing document. In other words, this theory brings a changed metaphor to citations: rather than being an indicator of the quality of the cited paper or a device to be used by the citing authors to defend their ideas, citation is a conversation between documents bound by linguistic rules (Cronin, 2001).

Based on this theory, citation and reference bear different meanings, given their distinct positions in the giver-receiver network (Cronin, 2000; D. J. Price, 1970; P. Wouters, 1998; P. F. Wouters & others, 1999). One version of this explanation is given by Cronin (2000) based on Pierce's sign triad (Gluck, 1997; which shows the signal vehicle [the signal itself], interpretant [the thing that translates a sign], and referent [the thing a sign denotes]), which is summarized in Table 3.1. For any reference that is embedded in a scientific paper, it points to both the bibliographic reference at the end of the paper as well as the document that is represented by the bibliographic reference. Its interpretant is the meaning of the sign-vehicle, which could be either located in the context of the reference, or in a large set of works by the same author. On the other hand, for a citation that is collected in Science Citation Index, they point to all the citing and cited works. Their meanings cannot be understood without the citation network between all these documents.

Signal vehicle	Interpretant	Referent
Embedded reference	Situated meaning	Work/object invoked
Citation in citation index	Connectedness and relatedness	Absent referent and other citing works

Table 3.1: The meanings of reference and citation in the symbolic theory

Based on these differences between references and citations, Cronin restated a criticism on the normative theory, which is that it ignores the "situated nature" (Cronin, 2000, p. 448) of references to use it to evaluate the academic merits of scientific outputs. On the other hand, the symbolic theory and the interpretive theory are different but largely compatible, given that one focuses on the linguistic meaning of the document, the other focuses on the position the document within microprocesses of scientific studies.

One type of studies that is directly inspired by this theory is to identify the entities embedded in the citations. A remarkable example is the concept of tiered citation proposed by Cronin (1994). In this paper, he argued that there are different levels of completeness in which a work is cited, from the whole body of papers to a specific concept. His work greatly extends our understandings of the inter-document relationship.

Another application of this theory is the question: to what extents do intentions of the authors or the readers determine the meanings of the citations/references. Henry Small (1978) and Blaise Cronin (2000) reached opposite conclusions by following the same theory: Small rooted the importance of author's intentions and vice versa. But both authors agreed that there are noisy and subjective elements from both ends of this relationship. The answer to this question has significant consequences on the analyses focusing on the motivation of citation, which started from Garfield's famous list of citation intentions (Garfield, 1965).

3.2 Relationship between the three models and their applicability to the studies of data objects

A quick glimpse of the literature is enough to draw the conclusion that each citation theory has its unique positions of the epistemology, ontology, methodology, and priority of citation studies, as well as the connection between documents (Luukkonen, 1997). The normative theory focuses on how the quantitative accumulation of citations by a document represents its quality that is directly connected to the academic reward system. The interpretive theory pays the closest attention to how citations are used by authors to defend their own claims against readers, as part of the overall knowledge production processes. The symbolic theory directly examines the relationship between symbolic meanings of the cited documents as represented in the citing works. All theories bind the nature of research questions, and collection and interpretation of data as much as they are bind by these factors.

Despite their differences, they are not totally mutual-exclusive. As mentioned above, the interpretive theory and the symbolic theory are seen by many as compatible with each other, even though they have distinct focuses. On the other hand, even though it is argued against by the other two theories, the normative theory is still playing a significant role in the landscape of scientometrics, and is still inspiring emerging studies. To build upon their similarities and connections, there have been efforts to integrate these paradigms to construct a unified theory of citation. A notable example of these efforts is from Susan Cozzens. In her literature review (Cozzens, 1981), she noted the possibilities of unifying these three approaches from a practical perspective: because all the theories are able to interpret, at least partly, scientists' actions. As a preliminary solution to this question, she noted the potentials of anthropological methods, especially the technique of discourse analysis discussed by Gilbert and Mulkay (1984), to serve as a bridge between all these theories. In a piece of her later work (Cozzens, 1989), she concluded that "citations should be seen primarily as rhetoric and only secondarily as recognition" (p. 445), based upon which a comprehensive quantitative model be established for citation data.

It is obvious that all these theories are contributing to the scholarship of cyberinfrastructure and can potentially contribute to the proposed study. It is true that the normative theory is no longer the focus of STS; however, it is still an important topic in information science and especially scientometrics. Linking to the topics of this literature review, many researchers have claimed that being integrated into the scholarly reward system is fundamental for such data objects to be given more attention across their full lifecycles, from its collection or production (Allen & Schmidt, 2014; Hettrick, 2016; Howison et al., 2015), archiving (Brody et al., 2007; Elman, Kapiszewski, & Vinuela, 2010), to (Kaye et al., 2009; Molloy, 2011; Reichman, Jones, & Schildhauer, 2011; Teuben et al., 2013). More studies are still needed to better understand how much impacts have been created by all kinds of digital objects.

From the perspectives of the latter two theories, it would be interesting to analyze as data objects become citable, what relationships are formed between them and the citing documents. A major method to answer this question is through content and context analysis, which will be discussed in the coming section.

3.3 Content and context analysis and its descendant

Content and context analysis is an important method in the construction of citation theories based on our discussions above. Even though the term of "content and context analysis" was not invented until the early 1980s (Small, 1982), this type of study, focusing on "particular message or statement within the citing document containing the reference" (Small, 1982, p. 288) can be traced back to the 1960s (Lipetz, 1965; Moravcsik & Murugesan, 1975). Content and Context analysis is normally composed of selecting a subset of scientific papers, manually coding the nature of the citations in these papers, and then developing a classification scheme of the different types of citations.

Cronin commented that the development of these classification schemes is not a "cumulative endeavor", even though there are some regularities between these schemes (Cronin, 1984, p. 35). His observation was supported by later works to identify and summarize the facets underlying these schemes, such as the one conducted by Zhang, Ding, & Milojević (2013), where the authors identified six principles of coding in the classification schemes they reviewed:

- Type of motivation
- Level of importance
- Type of resource
- Function of citing
- Type of disposition/sentiment
- Location of mentioning (p. 21)

It should be noted that just like all the classification schemes the authors reviewed, this scheme itself is subject to ambiguity that is part of the nature of the world. At least some of the categories cannot be separated distinctly, such as motivation and function. An example of such ambiguity in the list is Garfield's famous scheme of citation motivations (Garfield, 1965); items in this list are a good combination of motivation, function, and sentiment.

Despite these shortcomings, it is obvious that content and context analysis has important advantages for understanding the sociological and psychological factors as well as more detailed contexts behind citation behaviors. Even though most of the existing studies only classified the inter-document relationship based on one facet, as discussed by Zhang et al. (2013), it would be a beneficial effort to try a combination of the facets as well as to track various types of resources.

Another limitation of this method was the defined sample size, because it is time-consuming to conduct large-scale manual coding. However, the improved computational capacities and text techniques offer new possibilities to this stream of study. A significant example of this new approach is the series of studies conducted by a group of researchers at Cambridge University, where they adapted a scheme based on the work of Spiegel-Rösing (1977) and used machine learning techniques to classify citation functions based on language features around the citations (Teufel, Siddharthan, & Tidhar, 2006, 2009).

Our future studies will be making use of this approach to investigate more details about how data objects are cited in scientific publications. More specifically, the following questions will be pursued:

- If any scale of content and context classification can be applied to the citation/mention of digital objects?
- If there is any pattern of the citation of these objects, as compared with regular documents? If so, if there is any different pattern or relationship among different types of digital objects?
- If NLP or other text techniques can be applied to content and context analysis of digital objects?

3.4 From documents to entities

Traditionally, citation analysis is based on the links between documents only. Even as other type of objects are analyzed by citation studies, such as author, institution, country, journal, and knowledge domain (Ding et al., 2013), they simply serve as the level to which inter-document connections are aggregated to, rather than the research objects per se.

However, data objects are increasingly becoming the receiver of scientific citations. This phenomenon is due to some factors that have been reviewed in previous chapters. First, being able to be named is the prerequisite for such studies. Many citation studies make use of existing identifiers for data objects (Chao, 2011; Peters, Kraker, Lex, Gumpenberger, & Gorraiz, 2015, 2016); to the contrary, the lack of identifier for datasets but especially software makes it difficult to track these objects correctly (Li et al., 2016; Pan et al., 2015). Second, these new studies definitely benefit from data objects as new research front discussed above, and the new research questions, such as those proposed by Swanson (2015). For example, Meho (2007) discussed how the emergence of large-scale databases such as Web of Science and Scopus, more users are relying upon online interface to access to these databases, as well as new methods such as the PageRank algorithm, web citations, article-download counts, and h-index have changed citation analysis in significant ways. In terms of the new methods, Altmetrics is another important method that has broadly impacted the whole field of scientometrics (Piwowar, 2013;

Priem, 2013; Priem et al., 2010), which is especially important because it put a stronger focus on less official scientific communication means.

Ding and her colleagues summarized (2013) the variety of entities that are examined by scientometric studies by proposing the concept of "entitymetrics", which are categorized into the following three levels:

- Macro-level (evaluative) entities: author, journal, reference
- Meso-level (knowledge) entities: keywords
- Micro-level entities: dataset, method, biomedical entities

An important family of study in this broad topic is the examination of the citation patterns of these data objects as an indication of their impacts. Three themes of these studies can be identified. The first two, namely tracing the scientific impacts and demonstrating the benefits of sharing datasets, reflect the normative theory. The last theme, which is to prove the importance of a data citation standard, is connected to the fact that a data citation infrastructure is yet to be further improved and promoted.

The first category of these studies try to answer the question of what impacts have these data objects have created in scientific studies. Such examples include but not limit to Belter's study (2014) about how the three popular datasets archived in National Oceanographic Data Center are cited in other research outputs, the studies to trace the citation of all datasets in the Dryad Data Repository (He & Nahar, 2016; Mayo et al., 2016), Peters and her colleagues' study (2015) to compare the citation and altmetrics parameters on datasets, and a few efforts to rank the popularities of scientific software in scientific publication and public websites (Chao, 2011; Muenchen, 2012; Pan et al., 2015)

Quite similar for the first topic, another category of these studies focus on the benefits to the papers if the datasets are openly available (Dorch, 2012; Gleditsch, Metelits, & Strand, 2003; Henneken & Accomazzi, 2011; Ioannidis et al., 2009; Pienta, Alter, & Lyle, 2010; Piwowar et al., 2007; Piwowar & Vision, 2013). Even though these studies aim at the datasets that are embedded in scientific papers, the focus of these studies is still the impacts of the scientific papers per se. But these studies help to promote the importance of sharing the data objects.

The last theme in these studies is the necessity to develop and adopt uniformed citation standards for data objects, a topic that is more about the infrastructure than scholarship. This topic is addressed in many studies reviewed above, given the difficulties to identify the data objects in the scientific writings, even after using text-mining techniques. But it is also the major argument of some studies, which are beyond the scope of traditional scientometrics. Such examples included Piwowar, Carlson and Vision's study (2011) about the citation patterns of datasets from three data repositories (Gene Expression Omnibus, PANGAEA, and TreeBASE). A major conclusion from this study is that data accession number has its unique values despite the common use of DOI. Another example is the comparison of in-text data citation styles conducted by Mooney and

Newton (2012) where the authors identified the highly variant instructions and practices of data citation.

Our future studies will be conducted in the same area that has been established by all these works. But we will be extending them by tracing the relationship between different types of entities by making use of content and context analysis combined with textual techniques, such as NLP. Our aim is to have deeper understandings of these data objects and their related objects in the space of academic citation, which complement our studies discussed in Chapter two.

4 Conclusion

This literature review discusses a few topics around the positions of research data and scientific software in scientific practices from the traditions of STS and information science, and some potential paths forward based on existing studies. More specifically, we discuss the concept of cyberinfrastructure that my future studies will be located, especially its requirements and components that are relevant to data and software. In the second chapter, we review the concepts of data and software as well as their epistemological implications to scientific studies. After that, we discussed how data, software, scientific theories, and research methods could form relationship, in the light of the concept of package. In the last chapter, we talk about three major theories about citations and how citation analysis can be applied to digital objects that will be implemented in the future.

Based on these studies reviewed in this document, our future works will be focusing on using quantitative and qualitative methods to survey how data and software entities and their relationship are represented in scientific writings as compared with how they happen in the laboratory space. Both parts of this general research interest can be segmented into smaller questions, which are discussed across this literature review in corresponding sections. Moreover, in order to pursue the overall question, studies based on each tradition will be translated and compared, which is the ultimate goal of the proposed study.

REFERENCE

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.

Ahmed, S. F., Rodie, M., Jiang, J., & Sinnott, R. O. (2010). The European disorder of sex development registry: a virtual research environment. *Sexual Development*, *4*(4–5), 192–198.

Allen, A., & Schmidt, J. (2014). Looking before leaping: Creating a software registry. *ArXiv Preprint ArXiv:1407.5378*.

Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital preservation through archival collaboration: The data preservation alliance for the social sciences. *The American Archivist*, *72*(1), 170–184.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7), 16–07.

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure*. Retrieved from https://arizona.openrepository.com/arizona/handle/10150/106224

Baker, C. S. (2012). *Journal of heredity adopts joint data archiving policy*. Oxford University Press US. Retrieved from https://academic.oup.com/jhered/article-abstract/104/1/1/775539

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452. https://doi.org/10.1038/533452a

Ball, R., & Tunger, D. (2006). Science indicators revisited–Science Citation Index versus SCOPUS: A bibliometric comparison of both citation databases. *Information Services & Use*, *26*(4), 293–301.

Barga, R. S., Andrews, S., & Parastatidis, S. (2007). A virtual research environment (VRE) for bioscience researchers. In *Advanced Engineering Computing and Applications in Sciences, 2007. ADVCOMP 2007. International Conference on* (pp. 31–38). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/4401895/

Bates, M. J. (2005). Information and knowledge: an evolutionary framework for information science. *Information Research*, *10*(4), 10–4.

Beagrie, N., Beagrie, R., & Rowlands, I. (2009). Research data preservation and access: The views of researchers. *Ariadne*, (60). Retrieved from http://www.ariadne.ac.uk/issue60/beagrie-et-al/

Belter, C. W. (2014). Measuring the value of research data: a citation analysis of oceanographic data sets. *PloS One*, *9*(3), e92590.

Berman, F. (2008). Got data?: a guide to data preservation in the information age. *Communications of the ACM*, *51*(12), 50–56.

Berry, D. (2016). *The philosophy of software: Code and mediation in the digital age*. Springer. Retrieved from https://books.google.com/books?hl=en&lr=&id=GeYgDAAAQBAJ&oi=fnd&pg=PR1&dq=philosophy+of+software&ots=268PNRJJtr&sig=1eRNhPKVht_ZOlHqGt5RzKLb8og

Bertzky, M., & Stoll-Kleemann, S. (2009). Multi-level discrepancies with sharing data on protected areas: What we have and what we need for the global village. *Journal of Environmental Management*, *90*(1), 8–24. https://doi.org/10.1016/j.jenvman.2007.11.001

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Biology*, *12*(2), e1001797.

Bohle, S. (2013). What is E-science and How Should it be Managed. *Nature. Com, Spektrum Der Wissenschaft (Scientific American), Http://Www. Scilogs. Com/Scientific_and_medicallib Raries/What-Is-e-Science-and-How-Should-It-Be-Managed*.

Boisvert, R. F., Howe, S. E., & Kahaner, D. K. (1983). The GAMS classification scheme for mathematical and statistical software. *ACM SIGNUM Newsletter*, *18*(1), 10–18.

Borda, A., Careless, J., Dimitrova, M., Fraser, M., Frey, J., Hubbard, P., … Wiseman, N. (2006). Report of the working group on virtual research communities for the ost e-infrastructure steering group. Retrieved from https://eprints.soton.ac.uk/42074

Borek, L., Dombrowski, Q., Perkins, J., & Schöch, C. (2016). TaDiRAH: a Case Study in Pragmatic Classification. *Digital Humanities Quarterly*, *10*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/10/1/000235.html

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? Retrieved from https://works.bepress.com/borgman/238/

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078.

Borgman, C. L. (2015). *Big data, little data, no data: scholarship in the networked world*. MIT press. Retrieved from https://books.google.com/books?hl=en&lr=&id=gL8vBgAAQBAJ&oi=fnd&pg=PR7&dq=borgman+big+data+little+data&ots=I5a58Fenc3&sig=ohauGP8LpTh8VVnHG6gSFd2WXiw

Borgman, C. L., Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Ribes, D., … Calvert, S. (2013). Knowledge infrastructures: Intellectual frameworks and research challenges. *Deep Blue*. Retrieved from https://works.bepress.com/borgman/318/download/

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. Retrieved from http://works.bepress.com/furner/1/

Bowker, G. C. (2007). The past and the Internet. *Structures of Participation in Digital Culture*, 20–36.

Bowker, G. C., Baker, K., Millerand, F., & Ribes, D. (2009). Toward information infrastructure studies: Ways of knowing in a networked environment. In *International handbook of internet research* (pp. 97–117). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-9789-8_5

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press. Retrieved from https://books.google.com/books?hl=en&lr=&id=xHlP8WqzizYC&oi=fnd&pg=PR9&dq=bowker+sorting+things+out&ots=Mz8zrIt2pF&sig=fGLEpnQn5gFsyvQ-Pz3zqKjLAv0

Bowman, A. K., Crowther, C. V., Kirkham, R., & Pybus, J. (2010). A virtual research environment for the study of documents and manuscripts. Retrieved from https://ora.ox.ac.uk/objects/uuid:d7f250e0-9a95-4193-b476-8666ce5c3347

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brody, T., Carr, L., Gingras, Y., Hajjem, C., Harnad, S., & Swan, A. (2007). Incentivizing the open access research web: publication-archiving, data-archiving and scientometrics. *CTWatch Quarterly*, *3*(3).

Burton, A., & Treloar, A. (2009). Publish My Data: A composition of services from ANDS and ARCS. In *e-Science, 2009. e-Science'09. Fifth IEEE International Conference on* (pp. 164–170). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5380872

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, *22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. *Journal of the Association for Information Science and Technology*, *66*(9), 1747–1762.

Candela, L., Castelli, D., & Pagano, P. (2013). Virtual research environments: an overview and a research agenda. *Data Science Journal*, *12*, GRDI75–GRDI81.

Caplan, P. (2003). *Metadata fundamentals for all librarians*. American Library Association. Retrieved from https://books.google.com/books?hl=en&lr=&id=yt2863FismcC&oi=fnd&pg=PR5&dq=caplan+metadata+types&ots=AIaNLbB6I7&sig=35nTPHL7vH0bUxr0IHwInqMSYxA

CCSDS. (2002). Reference Model for an Open Archival Information System (OAIS). Retrieved from http://www.imaginar.org/taller/dppd/DPPD/46%20pp%20OAIS%20CCSDS.pdf

Chao, T. (2015). Mapping methods metadata for research data. *International Journal of Digital Curation*, *10*(1), 82–94.

Chao, T. C. (2011). Disciplinary reach: Investigating the impact of dataset reuse in the earth sciences. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–8. https://doi.org/10.1002/meet.2011.14504801125

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(15), 1.

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications*, *23*(3), 187–200.

Claerbout, J., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992* (Vols. 1–0, pp. 601–604). Society of Exploration Geophysicists. https://doi.org/10.1190/1.1822162

Clark, K. E. (1957). America's psychologists: A survey of a growing profession. Retrieved from http://psycnet.apa.org/psycinfo/2004-15427-000

Cole, J., & Cole, S. (1971). Measuring the Quality of Sociological Research: Problems in the Use of the" Science Citation Index". *The American Sociologist*, 23–29.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms*. MIT press. Retrieved from https://books.google.com/books?hl=en&lr=&id=aefUBQAAQBAJ&oi=fnd&pg=PR5&dq=introduction+to+algorithms+cormen&ots=dMbsTwYLiW&sig=NGDgcP7aDa4Rx3G0wnGi5QIRn4s

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, *1*(3), 193–203.

Costello, M. J. (2009). Motivating Online Publication of Data. *BioScience*, *59*(5), 418–427. https://doi.org/10.1525/bio.2009.59.5.9

Cozzens, S. E. (1981). Taking the measure of science: A review of citation theories. Retrieved from https://smartech.gatech.edu/handle/1853/32543

Cozzens, S. E. (1989). What do citations count? The rhetoric-first model. *Scientometrics*, *15*(5–6), 437–447.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023. https://doi.org/10.1098/rsta.2010.0165

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation*, *37*(1), 16–24.

Cronin, B. (1994). Brief Communication Tiered Citation and Measures of Document Similarity. *Journal of the American Society for Information Science (1986-1998); New York*, *45*(7), 537–538.

Cronin, B. (2000). Semiotics and evaluative bibliometrics. *Journal of Documentation*, *56*(4), 440–453.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. https://doi.org/10.1002/asi.1097

Da Costa, N. C., & French, S. (2003). *Science and partial truth: A unitary approach to models and scientific reasoning*. Oxford University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=snQSDAAAQBAJ&oi=fnd&pg=PP11&dq=science+and+partial+truth&ots=YoGfZOdx_A&sig=GBIETXbmyQv7obMpn5WhJXUAKx8

Daston, L. (2000). *Biographies of scientific objects*. University of Chicago Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=SsumCpb2QnAC&oi=fnd&pg=PR9&dq=biographies+of+scientific+objects&ots=DnuKWm-zzb&sig=n9HN6vGZXZD-YK2e9AoUla4jhfc

DataCite International Data Citation Metadata Working Group. (2015). *DataCite metadata schema for the publication and citation of research data version 3.1*. Retrieved from https://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

David, P. A. (1998). Common Agency Contracting and the Emergence of “Open Science” Institutions. *The American Economic Review*, *88*(2), 15–21.

Day, M. (1999). Metadata for digital preservation: an update. *Ariadne*, (22). Retrieved from http://www.ariadne.ac.uk/issue22/metadata

De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=ma4YjaKyM9cC&oi=fnd&pg=PR5&dq=Bibliometrics+and+Citation+Analysis+From+the+Science+Citation+Index+to+Cybermetrics&ots=1vZ2AV05zl&sig=x4G4j2k7GkL9f9MOoPRztXR_-MM

De Roure, D., & Goble, C. (2007). myExperiment–a web 2.0 virtual research environment. Retrieved from https://eprints.soton.ac.uk/263961

De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D., & Newman, D. (2008). myExperiment: Defining the social virtual research environment. In *eScience, 2008. eScience’08. IEEE Fourth International Conference on* (pp. 182–189). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/4736756/

Denis, J., & Goëta, S. (2014). Exploration, Extraction and ‘Rawification’. The Shaping of Transparency in the Back Rooms of Open Data. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2403069

Denis, J., & Goëta, S. (2017). Rawification and the careful generation of open government data. *Social Studies of Science*, 0306312717712473.

di Serafino, D., Maddalena, L., Messina, P., & Murli, A. (1998). Some perspectives on high-performance mathematical software. In *High Performance Algorithms and Software in Nonlinear Optimization* (pp. 1–23). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4613-3279-4_1

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the Impact of Entities. *PLOS ONE*, *8*(8), e71416. https://doi.org/10.1371/journal.pone.0071416

Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028

Dorch, S. B. F. (2012). On the citation advantage of linking to data: Astrophysics. *H-Prints and Humanities*. Retrieved from https://hal-hprints.archives-ouvertes.fr/hprints-00714715/document/

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 2053951716665128.

Drucker, J. (2011). Humanities Approaches to Graphical Display, *5*(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., … Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, *4*(3), 139–160. https://doi.org/10.1007/s12145-011-0083-6

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of Science*, *17*(2), 102–134.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Mit Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=K9_LsJBCqWMC&oi=fnd&pg=PR7&dq=paul+edwards+2010+infrastructure+knowledge&ots=EOBc07t1eV&sig=gMRU9IvGmSb89is2RFH6hZH5uCU

Edwards, P. N., Mayernik, M. S., Batcheller, A., Bowker, G., & Borgman, C. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 0306312711413314.

Egghe, L., & Rousseau, R. (1990). *Introduction to Informetrics : quantitative methods in library, documentation and information science*. Elsevier Science Publishers. Retrieved from http://eprints.rclis.org/6011/

Elman, C., Kapiszewski, D., & Vinuela, L. (2010). Qualitative data archiving: Rewards and challenges. *PS: Political Science & Politics*, *43*(1), 23–27.

Engeström, Y. (1990). *Learning, working and imagining: Twelve studies in activity theory*. Orienta-konsultit.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904.

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, *4*(5), e157.

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work (CSCW)*, *19*(3–4), 355–375.

Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. In *Opening science* (pp. 17–47). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-00026-8_2

Fenner, M. (2016). Announcing the Organization Identifier Project: a Way Forward [website]. Retrieved April 19, 2017, from https://blog.datacite.org/announcing-organization-identifier-project/

Field, D., Sansone, S.-A., Collis, A., Booth, T., Dukes, P., Gregurick, S. K., … Wilbanks, J. (2009). ’Omics Data Sharing. *Science (New York, N.Y.)*, *326*(5950), 234–236. https://doi.org/10.1126/science.1180598

Fleck, L. (1981). *Genesis and Development of a Scientific Fact*. University of Chicago Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=C50Jdn02wvMC&oi=fnd&pg=PP9&ots=1BVj_BdtFJ&sig=k7RsE1rcDshZqo0Y6F6930nUN44

Force, M. M., & Robinson, N. J. (2014). Encouraging data citation and discovery with the Data Citation Index. *Journal of Computer-Aided Molecular Design*, *28*(10), 1043–1048.

Foster, M. W., & Sharp, R. R. (2007). Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Reviews Genetics*, *8*(8), 633–639. https://doi.org/10.1038/nrg2124

Foucault, M. (2002). *The Order of Things: An Archaeology of the Human Sciences*. Psychology Press.

Fraassen, B. C. van. (2008). *Scientific Representation: Paradoxes of Perspective*. OUP Oxford.

Frické, M. (2015). Big data and its epistemology: Big Data and Its Epistemology. *Journal of the Association for Information Science and Technology*, *66*(4), 651–661. https://doi.org/10.1002/asi.23212

Froese, R., Lloris, D., & Opitz, S. (2004). The need to make scientific data publicly available: Concerns and possible solutions. *ACP-EU Fisheries Research Report*. Retrieved from http://www.vliz.be/en/imis?refid=209721

Fujimura, J. H. (1987). Constructing "Do-Able" Problems in Cancer Research: Articulating Alignment. *Social Studies of Science*, *17*(2), 257–293.

Fujimura, J. H. (1992). Crafting science: Standardized packages, boundary objects, and" translation.". *Science as Practice and Culture*, *168*, 168–169.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Garfield, E. (1965). Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings* (Vol. 1, pp. 189–92). National Bureau of Standards, Miscellaneous Publication 269, Washington, DC. Retrieved from https://books.google.com/books?hl=en&lr=&id=r56ZrbfTdkYC&oi=fnd&pg=PA189&dq=Can+Citation+Indexing+be+Automated%3F&ots=2vZrp8Qb6u&sig=XaDyDkpRe0BC3dP3cQdAAFkUMH8

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375.

Garfield, E. (1988). Derek Price and the Practical World of Scientometrics. *Science, Technology, & Human Values*, *13*(3/4), 349–350.

Garfield, E. (2004). The intended consequences of Robert K. Merton. *Scientometrics*, *60*(1), 51–61.

Garfield, E., & Merton, R. K. (1979). *Citation indexing: Its theory and application in science, technology, and humanities* (Vol. 8). Wiley New York. Retrieved from http://www.garfield.library.upenn.edu/cifwd.html

Garfield, E., & others. (1964). Science Citation Index-A new dimension in indexing. *Science*, *144*(3619), 649–654.

Gentleman, R. (2005). Reproducible research: a bioinformatics case study. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article2. https://doi.org/10.2202/1544-6115.1034

Gentleman, R., & Lang, D. T. (2004). Statistical Analyses and Reproducible Research. *Bioconductor Project Working Papers*. Retrieved from http://biostats.bepress.com/bioconductor/paper2

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, 113–122.

Gillespie, T. (2014). The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, *167*. Retrieved from http://books.google.com/books?hl=en&lr=&id=zeK2AgAAQBAJ&oi=fnd&pg=PA167&dq=info:jgo7uoqGxjUJ:scholar.google.com&ots=GngEQ-U0Ai&sig=QJknz9uFTfCu5bDut_SSypTc_6o

Gilliland-Swetland, A. J. (2000). Setting the state: Defining metadata. *Introduction to Metadata: Pathways to Digital Information. On-Line Document, Available at: Http://Www. Getty. Edu/Research/Institute/Standards/Intrmetadata/Index. Html*.

Gleditsch, N. P., Metelits, C., & Strand, H. (2003). Posting your data: will you be scooped or will you be famous. *International Studies Perspectives*, *4*(1), 89–97.

Gluck, M. (1997). Making sense of semiotics: privileging respondents in revealing contextual geographic syntactic and semantic codes. In *Proceedings of an international conference on Information seeking in context* (pp. 53–66). Taylor Graham Publishing. Retrieved from http://dl.acm.org/citation.cfm?id=267193

Gmür, M. (2003). Co-citation analysis and the search for invisible colleges: A methodological evaluation. *Scientometrics*, *57*(1), 27–57.

Gold, A. K. (2007). Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *Office of the Dean (Library)*, 16.

Grafton, A. (1997). *The footnote: a curious history*. Cambridge, Mass.: Harvard University Press.

Green, T. (2009). We need publishing standards for datasets and data tables. *Learned Publishing*, *22*(4), 325–327.

Greenberg, J. (2001). A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science and Technology*, *52*(11), 917–924.

Greenberg, J. (2003). Metadata and the world wide web. *Encyclopedia of Library and Information Science*, *3*, 1876–1888.

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=4hIQ5fGf-_oC&oi=fnd&pg=PA1&dq=hacking+representing+and+intervening&ots=5i8Is_HmJw&sig=2TTM_41cCOFVjKZmalHx5QJw2MM

Hammarberg, R. (1981). The cooked and the raw. *Information Scientist*, *3*(6), 261–267.

Harvard Information Infrastructure Project. (1995). *Public Access to the Internet*. MIT Press.

Harvey, D. R. (2010). *Digital curation: a how-to-do-it manual*. Neal-Schuman Publishers. Retrieved from http://www.bcin.ca/Interface/openbcin.cgi?submit=submit&Chinkey=424530

He, L., & Nahar, V. (2016). Reuse of scientific data in academic publications: An investigation of Dryad Digital Repository. *Aslib Journal of Information Management*, *68*(4), 478–494.

Henneken, E. A., & Accomazzi, A. (2011). Linking to Data - Effect on Citation Rates in Astronomy. *ArXiv:1111.3618 [Astro-Ph]*. Retrieved from http://arxiv.org/abs/1111.3618

Hettrick, S. (2016). Research software sustainability: Report on a Knowledge Exchange workshop.

Hey, T., Tansley, S., Tolle, K. M., & others. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA. Retrieved from https://www.fh-potsdam.de/fileadmin/user_upload/fb-informationswissenschaften/bilder/forschung/tagung/isi_2010/isi_programm/TonyHey_-__eScience_Potsdam__Mar2010____complete_.pdf

Hey, T., & Trefethen, A. E. (2002). The UK e-science core programme and the grid. *Future Generation Computer Systems*, *18*(8), 1017–1031.

Hong, N. C., Hole, B., & Moore, S. (2013). *Software papers: improving the reusability and sustainability of scientific software*. Technical Report 795303, WSSSPE1, 2013. http://dx. doi. org/10.6084/m9. figshare. 795303.

Howison, J., & Bullard, J. (2015). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155. https://doi.org/10.1002/asi.23538

Howison, J., Deelman, E., McLennan, M. J., Ferreira da Silva, R., & Herbsleb, J. D. (2015). Understanding the scientific software ecosystem and its impact: Current and future measures. *Research Evaluation*, *24*(4), 454–470. https://doi.org/10.1093/reseval/rvv014

IEEE Standards Coordinating Committee, & others. (1990). IEEE Standard Glossary of Software Engineering Terminology (IEEE Std 610.12-1990). Los Alamitos. *CA: IEEE Computer Society*.

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. https://doi.org/10.1038/nature10836

Ingersoll, R. C., Seastedt, T. R., & Hartman, M. (1997). A Model Information Management System for Ecological Research. *BioScience*, *47*(5), 310–316. https://doi.org/10.2307/1313192

Ingram, D. (1993). The Copernican Revolution revisited: paradigm, metaphor and incommensurability in the history of science-Blumenberg's response to Kuhn and Davidson. *History of the Human Sciences*, *6*(4), 11–35.

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., … van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155. https://doi.org/10.1038/ng.295

Kahin, B. (1993). Information technology and information infrastructure. *Lewis M. Branscomb (Hg.): Empowering Technology. Implementing a US Strategy. Cambridge/MA*, 135–166.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, *16*(3), 179–184.

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the notion of data curation in e-science: data managing and information infrastructuring in the long term ecological research (LTER) network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321–358.

Katz, D. S., Niemeyer, K. E., Smith, A. M., Anderson, W. L., Boettiger, C., Hinsen, K., … others. (2016). Software vs. data in the context of citation. *PeerJ Preprints*, *4*, e2630v1.

Katz, D. S., & Smith, A. M. (2015). Transitive Credit and JSON-LD. *Journal of Open Research Software*, *3*(1). Retrieved from http://openresearchsoftware.metajnl.com/articles/10.5334/jors.by/

Katz, M. J. (2009). *From research to manuscript: a guide to scientific writing*. Springer Science & Business Media.

Kaye, J., Heeney, C., Hawkins, N., De Vries, J., & Boddington, P. (2009). Data sharing in genomics—re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335.

Keller, E. F. (1995). *Refiguring Life: Metaphors of Twentieth-century Biology*. Columbia University Press.

Kelly, D. (2015). Scientific software development viewed as knowledge acquisition: Towards understanding the development of risk-averse scientific software. *Journal of Systems and Software*, *109*, 50–61.

Kelly, D., & Sanders, R. (2008). The challenge of testing scientific software. *CAST 2008: Beyond the Boundaries*, 30.

Kelly, D., Smith, S., & Meng, N. (2011). Software engineering for scientists. *Computing in Science & Engineering*, *13*(5), 7–11.

Kennedy, D. M. (2001). A primer on open source licensing legal issues: copyright, copyleft and copyfuture. . . *Louis U. Pub. L. Rev.*, *20*, 345.

King, G. (1995). Replication, Replication. *PS: Political Science and Politics*, *28*, 444–452.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE.

Kling, R. (1980). Social analyses of computing: Theoretical perspectives in recent empirical research. *ACM Computing Surveys (CSUR)*, *12*(1), 61–110.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., … others. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87–90). Retrieved from https://books.google.com/books?hl=en&lr=&id=Lgy3DAAAQBAJ&oi=fnd&pg=PA87&dq=jupyter+notebook&ots=N0G-7NqEeo&sig=LPKOEitsTSeZY5h7jAd9x-5d9U0

Knorr, K. D. (1981). The Manufacture of Knowledge An Essay on the Constructivist and Contextual Nature of Science. Retrieved from https://philpapers.org/rec/KNOTMO-2

Knorr-Cetina, K. (1992). *The couch, the cathedral, and the laboratory: On the relationship between experiment and laboratory in science*. Retrieved from https://kops.uni-konstanz.de/handle/123456789/11739

Knorr-Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press.

Knorr-Cetina, K. (2016). What if the Screens Went Black? The Coming of Software Agents. In *Beyond Interpretivism? New Encounters with Technology and Organization* (pp. 3–16). Springer, Cham. https://doi.org/10.1007/978-3-319-49733-4_1

Krumholz, H. M. (2014). Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Affairs*, *33*(7), 1163–1170. https://doi.org/10.1377/hlthaff.2014.0053

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, Ill.: University of Chicago press.

Lagoze, C., Lynch, C. A., & Daniel Jr, R. (1996). *The Warwick Framework: A Container Architecture for Aggregating Sets ofMetadata*. Cornell University. Retrieved from https://ecommons.cornell.edu/handle/1813/7248

Laine, C., Goodman, S. N., Griswold, M. E., & Sox, H. C. (2007). Reproducible research: moving toward research the public can really trust. *Annals of Internal Medicine*, *146*(6), 450–453.

Lambe, P. (2014). *Organising Knowledge: Taxonomies, Knowledge and Organisational Effectiveness*. Elsevier.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*, 70.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard university press. Retrieved from https://books.google.com/books?hl=en&lr=&id=sC4bk4DZXTQC&oi=fnd&pg=PA19&dq=latour+science+in+action&ots=WalGBrb7Nv&sig=0lO-95FniiSo3tGnq1hx697BG-o

Latour, B. (1990). Drawing things together. Retrieved from http://www.citeulike.org/group/10888/article/449517

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=HeptkrWIIpQC&oi=fnd&pg=PP2&ots=8EEDRnjozp&sig=xXQyyykXQDdwShk6Y6v3gc1aekg

Law, J. (1974). Theories and Methods in the Sociology of Science: An Interpretive Approach. *Social Science Information*, *13*(4–5), 163–172.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems Practice*, *5*(4), 379–393.

Law, J., & French, D. (1974). Normative and Interpretive Sociologies of Science. *The Sociological Review*, *22*(4), 581–595. https://doi.org/10.1111/j.1467-954X.1974.tb00509.x

Lawani, S. M., & Bayer, A. E. (1983). Validity of citation criteria for assessing the influence of scientific publications: New evidence with peer assessment. *Journal of the Association for Information Science and Technology*, *34*(1), 59–66.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, *6*(2), 4–37.

Leach, E. R. (1954). Political systems of highland Burma. A study of Kachin social structure. Retrieved from http://indianmedicine.eldoc.ub.rug.nl/root/L3/138l/

Leeuwen, T. (2005). Descriptive versus evaluative bibliometrics. *Handbook of Quantitative Science and Technology Research*, 373–388.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 1–3. https://doi.org/10.1016/j.shpsc.2011.10.001

Leonelli, S. (2014). What difference does quantity make? On the epistemology of Big Data in biology. *Big Data & Society*, *1*(1), 2053951714534395.

Leydesdorff, L. (1998). *Theories of Citation?* (SSRN Scholarly Paper No. ID 2279062). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2279062

Li, K., Greenberg, J., & Lin, X. (2016). Software Citation, Reuse and Metadata Considerations: An Exploratory Study Examining LAMMPS. In *Proceedings of* *the 79th ASIS&T Annual Meeting* (Vol. 53). Retrieved from http://dl.acm.org/citation.cfm?id=3017519

Lievrouw, L. A. (1989). The invisible college reconsidered: Bibliometrics and the development of scientific communication theory. *Communication Research*, *16*(5), 615–628.

Lipetz, B.-A. (1965). Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators. *American Documentation*, *16*(2), 81–90. https://doi.org/10.1002/asi.5090160207

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. In *Proceedings of the UK e-science All Hands meeting* (pp. 371–375). Retrieved from http://www.allhands.org.uk/2004/proceedings/papers/150.pdf

Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics*, *38*(1), 27–37.

Lynch, C. (1998). Identifiers and Their Role In Networked Information Applications. *Bulletin of the American Society for Information Science and Technology*, *24*(2), 17–20. https://doi.org/10.1002/bult.80

Lynch, C. (2006). Research Libraries Engage the Digital World: A US-UK Comparative Examination of Recent History and Future Prospects. *Ariadne*, (46). Retrieved from http://www.ariadne.ac.uk/issue46/lynch

Lynch, C. (2014). The next generation of challenges in the curation of scholarly data. *Research Data Management: Practical Strategies for Information Professionals. Purdue University Press, West Lafayette*, 395–408.

MacEwan, A., Angjeli, A., & Gatenby, J. (2013). The International Standard Name Identifier (ISNI): The Evolving Future of Name Authority Control. *Cataloging & Classification Quarterly*, *51*(1–3), 55–71. https://doi.org/10.1080/01639374.2012.730601

Maes, R., Rijsenbrij, D., Truijens, O., Goedvolk, H., & others. (2000). Redefining business: IT alignment through a unified framework. Retrieved from http://dare.uva.nl/ar/record/92240

Manovich, L. (1999). Database as symbolic form. *Convergence*, *5*(2), 80–99.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the Digital Humanities*, *2*, 460–475.

Marcus, A., & Menzies, T. (2010). Software is data too. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (pp. 229–232). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1882410

Marres, N. (2012). The redistribution of methods: on intervention in digital social research, broadly conceived. *The Sociological Review*, *60*(S1), 139–165.

Masterman, M. (1970). *The Nature of a Paradigm, w: Lakatos, I., Musgrave A.(eds.), Criticism and the Growth of Knowledge*. Cambridge University Press.

Mauthner, N. S., & Parry, O. (2009). Qualitative data preservation and sharing in the social sciences: On whose philosophical terms? *Australian Journal of Social Issues*, *44*(3), 291.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. Retrieved from https://books.google.com/books?hl=en&lr=&id=uy4lh-

WEhhIC&oi=fnd&pg=PP1&dq=mayer+schonberger+cukier&ots=Jsl5cjDRGS&sig=3wwRWAxYOPMZL04AeCbI7PCS6uI

Mayo, C., Vision, T. J., & Hull, E. A. (2016). The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *International Journal of Digital Curation*, *11*(1), 150–155.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legan and Political Sociology*, *1*, 115–126.

Merton, R. K. (1968). *Social theory and social structure*.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press. Retrieved from https://books.google.com/books?hl=en&lr=&id=zPvcHuUMEMwC&oi=fnd&pg=PR9&dq=sociology+of+science+merton&ots=x5TLRoi4vM&sig=VZEAlyOe4H0ohhxe7UTcMjLBzT8

Merton, R. K. (1977). *The Sociology of Science: An Episodic Memoir*. Southern Illinois University Press.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, *79*(4), 606–623.

Meyer, E. T., & Schroeder, R. (2015). *Knowledge machines: Digital transformations of the Sciences and Humanities*. MIT Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=QMnlBwAAQBAJ&oi=fnd&pg=PR5&dq=%22knowledge+machines%22&ots=O8ouDllwVq&sig=ezxXpadaAbbx5UGHHZ2aRd1ZZN0

Molloy, J. C. (2011). The Open Knowledge Foundation: Open Data Means Better Science. *PLOS Biol*, *9*(12), e1001195. https://doi.org/10.1371/journal.pbio.1001195

Mooney, H., & Newton, M. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication*, *1*(1). https://doi.org/10.7710/2162-3309.1035

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.

Muenchen, R. A. (2012). The popularity of data analysis software. *UR L Http://R4stats. Com/Popularity*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.565.3929&rep=rep1&type=pdf

Murray-Rust, P. (2008). Open Data in Science. *Nature Precedings*, (713). https://doi.org/10.1038/npre.2008.1526.1

Myers, C. R. (1970). Journal citations and scientific eminence in contemporary psychology. *American Psychologist*, *25*(11), 1041.

Myers, T., Trevathan, J., & Atkinson, I. (2012). The tropical data hub: a virtual research environment for tropical science knowledge and discovery. *International Journal of Sustainability Education*, *8*, 11–27.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Computer Horizons Washington, D. C. Retrieved from https://www.researchgate.net/profile/Francis_Narin/publication/284035800_Evalu ative_Bibliometrics_The_Use_of_Publication_and_Citation_Analysis_in_the_Evaluation_of_Scientific_Activity/links/565e188008aeafc2aac8d337.pdf

National Institutes of Health. (2003). NIH data sharing policy. *Retrieved From*.

National Research Council. (1999). *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. National Academies Press.

Nentwich, M. (2003). *Cyberscience: Research in the Age of the Internet*. Austrian Academy of Sciences Press Vienna. Retrieved from http://www.oeaw.ac.at/ita/en/publications/ita-books/cyberscience?sword_list%5B0%5D=nent

Nentwich, M., & König, R. (2012). *Cyberscience 2.0: Research in the age of digital social networks* (Vol. 11). Campus Verlag. Retrieved from https://books.google.com/books?hl=en&lr=&id=tAf4FhXWS0kC&oi=fnd&pg=PR5&dq=nentwich+cyberscience&ots=KCKCchjuId&sig=aF3d67CK-ghlOwBU6QeRcBFy0cI

Neuroth, H., Lohmeier, F., & Smith, K. M. (2011). TextGrid – Virtual Research Environment for the Humanities. *International Journal of Digital Curation*, *6*(2), 222–231. https://doi.org/10.2218/ijdc.v6i2.198

Neville, C. (2010). *The complete guide to referencing and avoiding plagiarism*. McGraw-Hill Education (UK). Retrieved from https://books.google.com/books?hl=en&lr=&id=dyBFBgAAQBAJ&oi=fnd&pg=PP1&dq=The+Complete+Guide+to+Referencing+and+Avoiding+Plagiarism&ots=IfSrmeJwvu&sig=CBoqdXkPJdsIK58OAW7N-UUICfQ

Niemeyer, K. E., Smith, A. M., & Katz, D. S. (2016). The challenge and promise of software citation for credit, identification, discovery, and reuse. *ArXiv Preprint ArXiv:1601.04734*. Retrieved from http://arxiv.org/abs/1601.04734

Noma, E. (1984). Co-citation analysis and the invisible college. *Journal of the Association for Information Science and Technology*, *35*(1), 29–33.

Norton, S., & Suppe, F. (2001). Why atmospheric modeling is good science. *Changing the Atmosphere: Expert Knowledge and Environmental Governance*, 67–105.

Norvig, P. (2008). *All we want are the facts, ma'am*. Retrieved 10/25/2013, from http://norvig. com/fact-check. html.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., … others. (2012). Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nature Methods*, *9*(4), 345–350.

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: A bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, *9*(4), 860–871.

Paskin, N. (2003). Components of drm systems identification and metadata. *Digital Rights Management*, 26–61.

Paskin, N. (2010). Digital object identifier (DOI) system. *Encyclopedia of Library and Information Sciences*, *3*, 1586–1592.

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, *12*(3), 30–32.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science (New York, N.Y.)*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Perkins, J., Dombrowski, Q., Borek, L., & Schöch, C. (2014). Building Bridges to the Future of a Distributed Network: From DiRT Categories to TaDiRAH, a Methods Taxonomy for Digital Humanities. *International Conference on Dublin Core and Metadata Applications*, 181–183.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research Data Explored: Citations versus Altmetrics. *ArXiv:1501.03342 [Cs]*. Retrieved from http://arxiv.org/abs/1501.03342

Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, *107*, 723–744. https://doi.org/10.1007/s11192-016-1887-4

Pienta, A. M., Alter, G. C., & Lyle, J. A. (2010). The enduring value of social science research: the use and reuse of primary research data. Retrieved from https://deepblue.lib.umich.edu/handle/2027.42/78307

Piwowar, H. A., Carlson, J. D., & Vision, T. J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–4.

Piwowar, H. A., & Chapman, W. W. (2008). A review of journal policies for sharing research data. In *ELPUB2008*. Retrieved from http://ocs.library.utoronto.ca/index.php/Elpub/2008/paper/view/684/0

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Price, D. de S. (1963). Big science, little science. *Columbia University, New York*, 119–119.

Price, D. J. (1970). Citation measures of hard science, soft science, technology, and nonscience. *Communication among Scientists and Engineers*, 3–22.

Prlić, A., & Procter, J. B. (2012). Ten Simple Rules for the Open Development of Scientific Software. *PLOS Computational Biology*, *8*(12), e1002802. https://doi.org/10.1371/journal.pcbi.1002802

Pröll, S., & Rauber, A. (2013). Scalable data citation in dynamic, large databases: Model and reference implementation. In *Big Data, 2013 IEEE International Conference on* (pp. 307–312). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6691588

Pröll, S., & Rauber, A. (2014). A scalable framework for dynamic data citation of arbitrary structured data. Retrieved from https://www.sba-research.org/wp-content/uploads/publications/Scalable%20Framework_paper.pdf

Purcell, A. (2014). Tool developed at CERN makes software citation easier. Retrieved from http://cds.cern.ch/record/1998637

Qin, J., Dobreski, B., & Brown, D. (2016). Metadata and Reproducibility: A Case Study of Gravitational Wave Research Data Management. *International Journal of Digital Curation*, *11*(1), 218–231.

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J., & Bussonnier, M. (2014). The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. In *AGU Fall Meeting Abstracts*. Retrieved from http://adsabs.harvard.edu/abs/2014AGUFM.H44D..07R

Rains, M. (2011). Creating a virtual research environment for archaeology. *Archaeology*, *2*, 159–170.

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705.

Rheinberger, H.-J. (1997). *Toward a History of Epistemic Things: Synthesizing Proteins in the Test Tube*. Stanford University Press.

Rheinberger, H.-J. (2005). A Reply to David Bloor: “Toward a Sociology of Epistemic Things.” *Perspectives on Science*, *13*(3), 406–410.

Rice, J. R. (2013). Mathematical Software. In *Encyclopedia of Computer Science* (pp. 1093–1096). Chichester, UK: John Wiley and Sons Ltd. Retrieved from http://dl.acm.org/citation.cfm?id=1074100.1074579

Rosenberg, D. (2013). *Data before the fact*. Raw data” is an oxymoron. Cambridge, Mass.: MIT Press. Retrieved from http://pages.uoregon.edu/koopman/courses_readings/colt607/rosenberg_data-before-fact_proofs.pdf

Rushby, N. (2015). Editorial: Data papers. *British Journal of Educational Technology*, *46*(5), 899–903. https://doi.org/10.1111/bjet.12337

Sahoo, S. S., Sheth, A., & Henson, C. (2008). Semantic provenance for escience: Managing the deluge of scientific data. *IEEE Internet Computing*, *12*(4). Retrieved from http://ieeexplore.ieee.org/abstract/document/4557978/

Sarwar, M. S., Doherty, T., Watt, J., & Sinnott, R. O. (2013). Towards a virtual research environment for language and literature researchers. *Future Generation Computer Systems*, *29*(2), 549–559.

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PloS One*, *4*(9), e7078.

Sawyer, S. (2008). Data Wealth, Data Poverty, Science and Cyberinfrastructure [1]. *Prometheus*, *26*(4), 355–371. https://doi.org/10.1080/08109020802459348

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities*, *2*(3), 2–13.

Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., … Rosenthal, N. (2009). Post-publication sharing of data and tools. *Nature*, *461*(7261), 171–173. https://doi.org/10.1038/461171a

Sharma, N. (2004). The origin of DIKW Hierarchy. *Go. Webassistant. Com*, *11*. Retrieved from https://erealityhome.wordpress.com/2008/03/09/the-origin-of-dikw-hierarchy/

Sinnott, R. O., & Stell, A. J. (2011). Towards a Virtual Research Environment for International Adrenal Cancer Research. *Procedia Computer Science*, *4*, 1109–1118. https://doi.org/10.1016/j.procs.2011.04.118

Slota, S. C., & Bowker, G. C. (2016). How Infrastructures Matter by Stephen C. Slota and Geoffrey C. Bowker. In *The Handbook of Science and Technology Studies* (Fourth edition; Amazon version). Boston: MIT Press. Retrieved from https://mit-press.myshopify.com/products/chapter-18-how-infrastructures-matter

Small, H. (1982). Citation context analysis. *Progress in Communication Sciences*, *3*, 287–310.

Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics*, *60*(1), 71–79. https://doi.org/10.1023/B:SCIE.0000027310.68393.bc

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*(3), 327–340.

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86. https://doi.org/10.7717/peerj-cs.86

Smith, L. C. (1981). Citation analysis. *Library Trends*, *30*(1), 83–106.

Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, 97–113.

Star, S. L. (1985). Scientific work and uncertainty. *Social Studies of Science*, *15*(3), 391–427.

Star, S. L. (1989). *Regions of the Mind: Brain Research and the Quest for Scientific Certainty*. Stanford University Press.

Star, S. L. (1995). *Ecologies of knowledge: Work and politics in science and technology*. SUNY Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=wpv9HZKaCnwC&oi=fnd&pg=PR9&dq=%22ecologies+of+knowledge%22+star&ots=nSMvvWcPNB&sig=icLdrjyYOpuklVL736ip-hL1e4w

Star, S. L., & Ruhleder, K. (1994). Steps Towards an Ecology of Infrastructure: Complex Problems in Design and Access for Large-scale Collaborative Systems. In

Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (pp. 253–264). New York, NY, USA: ACM. https://doi.org/10.1145/192844.193021

Starr, J., & Gastl, A. (2011). isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, *17*(1), 9.

Steiner, C. M., Agosti, M., Sweetnam, M. S., Hillemann, E.-C., Orio, N., Ponchia, C., … others. (2014). Evaluating a digital humanities research environment: the CULTURA approach. *International Journal on Digital Libraries*, *15*(1), 53.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Swanson, D. R. (2015). Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18. https://doi.org/10.1353/pbm.1986.0087

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Teuben, P., Allen, A., Berriman, B., DuPrie, K., Hanisch, R. J., Mink, J., … Taylor, M. (2013). Ideas for Advancing Code Sharing (A Different Kind of Hack Day). *ArXiv Preprint ArXiv:1312.7352*.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1610091

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80–87). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1654612

Unsworth, J. (2006). *Our Cultural Commonwealth: the report of the American Council of learned societies commission on cyberinfrastructure for the humanities and social sciences*. ACLS: New York,. Retrieved from https://www.ideals.illinois.edu/handle/2142/189

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society*, *12*(2), 197.

Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative knowledge work and practices of trust: sharing environmental planning data sets. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work* (pp. 335–343). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=289508

Van Raan, A. F. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Vaughan, K., Hayes, B. E., Lerner, R. C., McElfresh, K. R., Pavlech, L., Romito, D., … Morris, E. N. (2013). Development of the research lifecycle model for library services. *Journal of the Medical Library Association : JMLA*, *101*(4), 310–314. https://doi.org/10.3163/1536-5050.101.4.013

Virgo, J. A. (1977). A statistical procedure for evaluating the importance of scientific papers. *The Library Quarterly*, 415–430.

Vision, T. J. (2010). Open Data and the Social Contract of Scientific Publishing. *BioScience*, *60*(5), 330–331. https://doi.org/10.1525/bio.2010.60.5.2

Vogel, R. L. (1998). Why scientists have not been writing metadata. *Eos, Transactions American Geophysical Union*, *79*(31), 373–380. https://doi.org/10.1029/98EO00284

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *ArXiv Preprint ArXiv:1309.5821*. Retrieved from https://arxiv.org/abs/1309.5821

Wickham, H. (2015). *R packages*. O’Reilly Media, Inc. Retrieved from https://books.google.com/books?hl=en&lr=&id=DqSxBwAAQBAJ&oi=fnd&pg=PR3&dq=r+packages+wickham&ots=am14LUQFHb&sig=3eFJlkvBBo3lHRqjSPlkPgMOFWc

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

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121–136.

Wirth, N. (1978). *Algorithms+ data structures= programs*. Prentice Hall PTR. Retrieved from http://dl.acm.org/citation.cfm?id=540029

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93–103.

Woolgar, S. (1982). Laboratory studies: A comment on the state of the art. *Social Studies of Science*, *12*(4), 481–498.

Wouters, P. (1998). The signs of science. *Scientometrics*, *41*(1–2), 225–241.

Wouters, P. F., & others. (1999). The citation culture. Retrieved from http://dare.uva.nl/document/2/8218

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. *IJDC*, *6*(1), 214–225.

Zeleny, M. (1987). Management support systems: towards integrated knowledge management. *Human Systems Management*, *7*(1), 59–70.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490–1503.

Zimmerman, A. (2007). Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries*, *7*(1–2), 5–16.

Zimmerman, A. S. (2008). New knowledge from old data the role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631–652.

Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. *Scientometrics*, *12*(5–6), 329–338.