Chelsea Crane

Dr. P.A.F. Verhaar (MA – thesis)

Dr. Kristina Hettne (second reader)

17 February 2020

Word Count: 15259

A Critical Assessment of The FAIR Guiding Principles in Book History

Table of Contents

Introduction

Humanities researchers are increasingly adopting methods and tools traditionally found in the Sciences and Social Sciences to analyze quantitative data about their research.[1] Digital research data has become larger and more complex not only within the Social Sciences, but within the Humanities as well, leading to the development of new tools and computational methods to find, interpret and ultimately publish data online.[2]

Recently, researchers have begun to discuss ways to 'improve the infrastructure supporting the reuse of scholarly data,'[3] not only by making it easier to find data in the first place, but to also access, reuse, and link to other relevant datasets, thereby enriching and ensuring the long-term usability and sustainability of such data. In 2014, a group of researchers from a variety of academic disciplines came together at the Lorentz Center in Leiden, The Netherlands to discuss and eventually publish a set of guiding principles for the findability, accessibility, interoperability, and reusability of data, which they ultimately coined as the FAIR Guiding Principles.[4] Since the publication of the principles in 2016, FAIR has been gaining momentum among researchers and stakeholders within the social sciences, together with governing bodies, funding agencies, and markedly the European Commission Directorate-General for Research and Innovation.[5]

[1] Leo Lahti, Niko Ilomaki, Mikko Tolonen, 'A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800', *Liber Quarterly*, 2 (2015), pp. 87-116, <10.18352/lq.10112> (2 November, 2019).

[2] M. D. Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific Data,* 3 (2016), pp. 160018 <doi: 10.1038/sdata.2016.18 (2016)> (30 October 2019).

[3] Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', p. 1.

[4] Barend Mons, 'FAIR Data Publishing Group', *FORCE11*, n.pag. <https://www.force11.org/group/fairgroup> (2 November 2019).

[5] Anon., 'H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020', *European Commission Directorate-General for Research & Innovation*, July 2016, pp.1-12 <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf> (2 November 2019).

In the following chapters, we examine in detail the FAIR Guiding Principles, what they are, why they were created, as well as highlight the difference between Linked Open Data and FAIR. We then give an overview of the current environment of humanities scholarship, paying particular attention to the Digital Humanities and book history, as well as any current applications of FAIR within these fields. We explore some of the reasons that the implementation of FAIR is significantly slower in the Humanities compared to the Sciences or Social Sciences by highlighting some of the challenges faced by humanities scholars in terms of producing and quantifying digital research data that is also easily findable and reusable, while taking time to discuss issues found in all disciplines such as IP, copyright, and privacy laws, as well as issues concerning authenticity, authority, trust, verification, and uncertainty relevant to open-source platforms and digital assets. A case study is then presented using a database that was created using information from the original book catalogue and cashbooks from the Bibliotheca Thysiana, a seventeenth-century library located in Leiden, The Netherlands. After analyzing the quality of the data from the Thysiana based on the requirements of the FAIR Principles, we then utilize the steps in the FAIRification Process by applying each to the database one at a time, highlighting any challenges along the way. Finally, we conclude with thoughts and criticisms on the feasibility of the application of FAIR onto a humanities database, while speculating if it is indeed a guide that can be implemented practically in the field of book history and if there is any truth to the fear of "pigeon-holing" researchers through strict frameworks, finishing with how we see FAIR working in the future.

Chapter 1: FAIR at a Glance

1.1

What Is FAIR?

Since the 2016 *Scientific Data* publication of the 'FAIR Guiding Principles for scientific data management and stewardship,' the FAIR Principles have been widely adopted and championed as a viable solution to the ongoing issues of reusability of valuable scholarly research data.[6] With the early adoption by research initiatives specifically within the European Union, including the new EU Framework Program Horizon2020, the FAIR Principles appear to be on their way to becoming the foundation on which research policy and data management plans are created.[7] The creators of the FAIR Principles correctly identified the lack of a cohesive set of standards across all disciplines for the publishing of digital scholarly data, with far too many computational methods, analytical workflows, and tools created for specific projects getting in the way of the discovery, evaluation, use and reuse of digital research objects.[8] This sentiment is shared by many data producers and researchers, with some arguing that the sharp increase over the past decade of 'arbitrarily different, incompatible standards' have done nothing but increase the rate at which research communities have divided and fragmented amongst themselves, furthering the need for a sense of cohesiveness.[9] Further, with the 'rapidly growing and evolving data environment,' novel technologies and 'new, more complex data-types' currently under development, including the rise of general-

[6] B. Mons, et al, 'Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud,' *Information Services & Use*, vol. 37, no. 1, Jan. 2017, pp. 49–56. *content.iospress.com* <doi:10.3233/ISU-170824> (1 November 2019).

[7] M. Boeckhout, et al, 'The FAIR Guiding Principles for Data Stewardship: Fair Enough?' *European Journal of Human Genetics*, vol. 26, no. 7, July 2018, pp. 931–36. *www.nature.com*, <doi:10.1038/s41431-018-0160-0> (1 November 2019).

[8] Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', p. 1.

[9] S.A., Sansone, et al., 'FAIRsharing as a Community Approach to Standards, Repositories and Policies', *Nature Biotechnology*, vol. 37, no. 4, Apr. 2019, pp. 358–67. *www.nature.com* <doi:10.1038/s41587-019-0080-8> (30 October 2019).

purpose repositories in which the data-types confronted by users may be unpredictable and random, machine-actionability is an increasingly pertinent issue which demands our attention.[10]

The authors of the FAIR Principles highlight the fact that unlike other contributors to the debate on reusability and Open Science, their principles focus not only on the human researcher but equally on the machines they use to conduct their research, since 'interoperability technologies and standards at the data/repository level' which aid machines in efficient data discovery and integration is a 'first-priority for good data stewardship.'[11] In the following section, we will take an in-depth look at each of the individual FAIR Principles, the definitions of which have been summarized directly from the *GO FAIR* website, while keeping in mind that they have been created as guidelines, not standards, and therefore are intentionally open-ended and ambiguous so as to be more easily adaptable.

1.2

The FAIR Guiding Principles[12]

FAIR is an acronym that stands for Findable, Accessible, Interoperable, and Reusable, representing the four main pillars of the principles. Although the principles are indeed linked, they can be implemented separately and independently from each other. The FAIR Principles are intended to be used not as a standard but as a guide that precedes implementation, the elements of which can be broken down into fifteen measurable and actionable steps:[13]

[10] Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', p. 4.
[11] Ibid.
[12] All principles summarized from the *GO FAIR* website; Anon., 'FAIR Principles', *GO FAIR,* n.pag. <https://www.go-fair.org/fair-principles/> (30 October 2019).
[13] Wilkinson, et al., p. 5.

Findable –

F1. ‘(Meta)data are assigned globally unique and persistent identifiers’

- This is considered to be the most important principle as each subsequent step relies upon obtaining unique and persistent identifiers (PIDs), which when applied to each instance of data and metadata ensures that the reference to these data continues to be reliable. PIDs not only assist human researchers with finding, reusing, and citing data, but they also aid computers in the same activities in addition to automatically interpret and integrate the data they find. There are two conditions which must be met for identifiers, the first being that they are globally unique, and second, that they are persistent. This ensures that there is no chance of reassigning or replicating an identifier to an already existing one without reference to the original, while also safeguarding the web links from becoming inactive.

F2. ‘Metadata are described with rich metadata’

- This principle describes the importance of including rich, detailed metadata to each instance of data within a given project. The point here is that even without identifiers the data should still be findable because it contains rich metadata, both intrinsic (data captured automatically at the time of data creation, such as date and timestamps) and contextual metadata (data describing what it is, how it was created, by whom, physical descriptors etc.). Ensuring that each data instance is coupled with robust metadata helps with the locating, reusing, and citing of the data, and is considered to be the next most important step after the application of persistent identifiers.

F3 – ‘Metadata clearly and explicitly include the identifier of the data they describe’

- The relationship between the dataset and the metadata should be explicitly stated by clearly including the dataset’s ‘globally unique and persistent identifiers’ in the metadata, as the two usually come in separate files, thereby ensuring that both the data and metadata are considered together as findable.

F4 – '(Meta)data are registered or indexed in a searchable resource'

- Rich metadata and unique identifiers are not enough to guarantee that data is actually findable, as 'perfectly good data resources may go unused simply because no one knows they exist'. Ensuring the findability of valuable research (meta)data by both machines and humans necessitates indexing or registering (meta)data in searchable resources and repositories so they remain findable.

A1 – '(Meta)data are retrievable by their identifier using a standardized communication protocol'

- Data retrieval must be made possible through the use of both tools and communication methods that are widely-used and unspecialized. Explicitly state how and by whom the data can be accessed. Avoiding the use of methods and protocols that have 'limited implementations, poor documentation, and components involving human intervention' is equally pertinent to ensuring accessibility.

A1.1 – 'The protocol is open, free and universally implementable'

- In general, accessibility should be possible by anyone 'with a computer and an internet connection,' and at minimum access to the metadata should be done through 'free (no-cost) and open (-sourced)' protocols, ensuring that they are commonly useable for conducting data retrieval.

A1.2 – 'The protocol allows for an authentication and authorization where necessary'

- Providing the 'exact conditions under which data are accessible' is key in order to ensure that machines are able to automatically detect any requirements for accessing restricted data and either execute the necessary steps or notify the user on how to do so. This may involve the creation of a user account in order to access data held in a repository, with requirements for 'user-specific rights' for each dataset.

A2 – 'Metadata should be accessible even when the data is no longer available'

- Even if the weblinks associated with data become inactive, the metadata should continue to persist. It is not uncommon for datasets to degrade or disappear entirely since maintaining their online existence costs money and time, therefore it is advised that the metadata remain available as this is most useful in finding the right institutions, people or publications when the original data has been lost. This principle is related to indexing issues in F4, discussed below.

I1 – '(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation'

- This principle not only refers to the exchange and interpretation of human languages (a.k.a. be clear and do not use "dead" languages), but machine language as well. Interoperability between machines means that computer systems have at minimum a basic understanding of each other's data exchange formats, without the use of bespoke algorithms, specialized translators, or mappings. Guaranteeing ideal 'automatic findability and interoperability of datasets' means using generic, controlled vocabularies, ontologies and thesauri (with resolvable, globally unique persistent identifiers), as well as a sound data model which structures and provides sufficient framing for the (meta)data.

I2 – '(Meta)data use vocabularies that follow the FAIR principles'

- In terms of the metadata describing datasets, the employed vocabulary must not only be 'controlled' (i.e. standardized) but also documented and resolvable with the use of globally persistent and unique identifiers. Ensuring findability, this documentation must be accessible by anyone using the dataset.

I3 – '(Meta)data include qualified references to other (meta)data'

- It is necessary to be explicit when cross-referencing or referring to other data, as the goal here is to create and maintain as many significant links between (meta)data assets as is possible, thereby 'enriching the contextual knowledge about the data' while also balancing the 'time/energy' necessary to create a sound data model. The authors use the example that the statement '*X is a regulator of Y'* is more useful than '*X is associated with Y*' or '*X see also Y*', and explain that it is important to refer to scientific links between datasets as well as any additional or complementary datasets which may build upon, or be needed to complete the data (including proper citation of all datasets and associated PIDs).

R1 – '(Meta)data are richly described with a plurality of accurate and relevant attributes'

- Related to principle F2, this principle states the importance of including as much detailed metadata as possible, whether or not it seems "relevant" to the publisher, thus providing necessary information on how to find the dataset in addition to richly describing the dataset both contextually and intrinsically. 'Plurality' is key here and it is emphasized that the publisher should not try to anticipate the data consumer's needs or identity when generating metadata, but to instead be as 'generous as possible' and provide metadata that can be parsed and evaluated for relevancy by both humans and machines.

R 1.1 – "(Meta)data are released with a clear and accessible usage license'

- This principle deals with '*legal* interoperability,' as opposed to *'technical* interoperability' as is outlined under the 'I' principles. Licensing restrictions complicate automated searches, so it is imperative that usage rights are clearly stated in the metadata because failing to do so will lessen the chance that the dataset will be reused by organizations or institutions that cannot comply with the licensing restrictions. This information should be explicitly stated and retrievable by both humans and machines.

R 1.2 – "(Meta)data are associated with detailed provenance'

- Building upon previous principles, the intention behind R 1.2 is to clearly state the origin of the data, including its full workflow from who created it and how, if it has been processed, transformed, previously published or completed in any way, and who to acknowledge or cite. This information is preferably published using machine-readable formatting.

R 1.3 – '(Meta)data meet domain-relevant community standards'

- If a community standard or method of best practice is available, then it should be employed. Ideally, data publishers should follow a common template by which data is organized in a standardized way, utilizing commonly used and sustainable file formats, vocabulary, data-types, and documentation (i.e. metadata). This ensures that (meta)data is published with the highest chances of its 'use(ability) for the community', which is the 'primary objective of FAIRness'. Although there may be reasons why a data publisher must use methods outside of community-standard best practices, these reasons must nonetheless be explicitly stated in the metadata.

1.3

FAIR IS…and is NOT

In recent years, scholars, researchers, and stakeholders from fields ranging from the social sciences, STEM, and increasingly the humanities, have come together to collaborate on finding a solution for issues regarding the promotion of Open Science and reusable research data.[14] The creation of the FAIR Guiding Principles has undoubtedly been one of the most successful of these collaborations at least in terms of widespread adoption, seen especially within the European Union, such as by the Directorate General for Research and Innovation of the European Commission and Science Europe.[15] As the authors of the FAIR Principles suggest, the main reason for why it takes 'several weeks (or months)' for specialists to collect data is not due to the lack of suitable technology, but is instead that 'we do not pay our valuable digital objects the careful attention they deserve when we create and preserve them.'[16] The current obstacles surrounding the discovery and reusability of research data is a major problem for data consumers, stakeholders, and researchers, and the continued lack of a 'minimal set of community-agreed guiding principles and practices' prevents both humans and machines from efficiently locating, integrating, and citing data that will otherwise go undiscovered.[17]

The authors of FAIR take great care to emphasize the importance of achieving 'FAIRness' for both 'human-driven and machine-driven activities,' a fact that they claim 'distinguishes them from many peer initiatives' contributing to the debate.[18] Their reasoning behind placing so much focus on machine-actionability has to do with the fact that unlike machines, humans are limited in their ability to 'operate at the scope, scale, and speed necessitated by the scale of contemporary scientific data and complexity of e-Science,' thereby requiring human scholars to increasingly rely upon machines for data discovery and integration.[19] Although humans are traditionally more adept at identifying the 'semantics' or underlying intent behind a given digital asset, machines are much better equipped to tackle the

[14] Sansone, et al., 'FAIRsharing as a Community Approach to Standards, Repositories and Policies', p 1.

[15] Mons, et al, 'Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud', p. 49.

[16] Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', p. 3.

[17] Ibid.

[18] Ibid.

[19] Ibid.

large-scale computational procedures necessary in our current quickly-evolving digital environment.[20] As they point out, our very lack of speed-driven, complex computational abilities 'necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols' that they will come up against as they guide themselves through the 'global data ecosystem.'[21] This also helps to explain the undeniable importance the authors have placed on the need for rich, detailed metadata as the machines must be capable of finding and keeping track of a digital objects' provenance, thereby ensuring that it can be properly cited.[22]

One of the more appealing characteristics of the FAIR Principles is that they provide an extensive, but not necessarily exhaustive, guideline of actionable steps which researchers can implement into their own data research projects without having to prescribe to a restricting standard. However, the authors of FAIR noted that as widespread adoption and support for the principles grew, so did the number of interpretations, some of which took varying liberties with their original intention.[23] This realization eventually encouraged a few of the authors to redefine exactly what 'FAIRness is, and is not.'[24]

As the authors state, when talking about FAIRness, they are indeed referring to a 'set of principles' which focus quite exclusively on guaranteeing the use and reuse of digital research objects, thereby ensuring their sustainability and value within research communities.[25] The FAIR Principles offer suggestions as to how they may be implemented in a practical sense, but they intentionally refrain from becoming prescriptive, especially in terms of technical requirements (for example, using specific types of software or tools, RDF, or other Semantic Web frameworks

[20] Ibid.
[21] Ibid.
[22] Ibid.
[23] Mons, et al, 'Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud', p. 49.
[24] Ibid.
[25] Ibid., p. 50.

and technologies).[26] Even still, despite the fact that it was explicitly stated in the original paper that the FAIR Guiding Principles are *not* a standard, since its publication they have been repeatedly cited as such.[27] The issue with this is that researchers or stakeholders interested in implementing FAIR into their own projects may feel held back by their perceived lack of appropriate technologies, tools, or expertise, when in reality the FAIR Principles can be applied incrementally and in almost any order according to the abilities and resources available to the researcher.[28]

Perhaps the greatest misconception of the principles is that "FAIR" automatically means "Open." Though the authors concede that the "A" in FAIR does represent Accessibility, this in no way implies that all data must be fully open and available to anyone in order for it to qualify as FAIR.[29] The reasons behind implementing data restriction are valid, including privacy protection (especially when dealing with public health data), copyright, and proprietary information, and it has been suggested that there is still considerable work to be done in terms of identifying the most 'responsible ways of facilitating data sharing.'[30] The difficulties surrounding data sharing cannot be easily resolved by one set of guiding principles, and the authors themselves note that they specifically do not 'address moral and ethical issues pertaining to the openness of data.'[31] This is why it is so imperative that all FAIR research data contain extensive and accessible metadata so that in the case of a user attempting to access "closed" or restricted data, the metadata can direct them either to the organization which owns the data or through the process of accessing the data themselves.[32] The authors believe that although data need not be

[26] Ibid., pp. 50-51.
[27] Ibid., p. 51.
[28] Ibid., pp. 52-53.
[29] Ibid., p. 51.
[30] Boeckhout, et al, 'The FAIR Guiding Principles for Data Stewardship: Fair Enough?'

[31] Mons, et al, 'Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud', p. 51.
[32] Ibid.

freely available or open in order to be FAIR, data should nevertheless come with explicit instructions for its access and reuse, stating:

> 'transparent but controlled accessibility of data and services, as opposed to the ambiguous blanket-concept of "open", allows the participation of a broad range of sectors – public and private – as well as genuine equal partnership with stakeholders in all societies around the world.'[33]

Funding is one of the issues that continues to perplex research scholars and organizations in terms of promoting a balanced and fair system of Open Science.[34] Barend Mons, one of the co-authors of the FAIR Principles, has recently suggested that a viable solution to the current 'imbalance' that exists between developing and developed nations that are publishing research data may lie in the "closed" paying for the "open".[35] This seemingly simply solution does possess some clout, since it has become a real problem for developing nations to pay the often exorbitant publication fees required of scholarly journals, not to mention that they are also usually dealing with poor Internet access and inadequate research funding as well.[36] Mons proposes that 'only those who wish to keep research discoveries private, pay,' whereas everyone else has 'free authorship and copyright if they are prepared to share their knowledge without restrictions.'[37] Mons believes that this fundamental shift in the way in which research is currently being funded would lead to 'millions' more scientists from developing nations the ability to participate in advancing research and lend a much needed 'boost' to Open Science, as well as the creation of a welcome bias towards more efficient sharing of scientific discovery.[38]

[33] Ibid., p. 52.
[34] B. Mons, 'When privacy-bound research pays for open science,' *EuroScientist Journal*, n.pag. April 2016, <https://www.euroscientist.com/privacy-bound-research-pays-open-science/> (2 November 2019).
[35] Ibid.
[36] Ibid.
[37] Ibid.
[38] Ibid.

It may be pertinent at this point to dive a bit deeper into the ways in which Linked Open Data (LOD) and FAIR blend, as well as into the ways in which they deviate from one another. Researchers Ali Hasnain and Dietrich Rebholz-Schuhmann explore the relationship between Sir Tim Berners-Lee's LOD 5-Star Principles[39] and FAIR, specifically in terms of whether or not the FAIR Principles reuse LOD Principles and if so, to what extent they overlap or build upon them.[40] Created in 2010, the LOD 5-Star scheme follows a system whereby the more open, accessible, and inter-linked the data, the more "stars" are gained.[41] Berners-Lee describes Linked Open Data as data which is published on the Semantic Web under an open license that does not obstruct its reuse as free data, in contrast to simply Linked Data which is data that is similarly linked on the Semantic Web, but is published under a restricted license.[42]

Hasnain and Rebholz-Schuhmann assert that both LOD and FAIR refrain from prescribing specific implementation choices, technologies, and tools, and both eschew the label of "standard".[43] Yet while the LOD 5-Star system was specifically created for Open Data in order to make data more accessible and reusable, the FAIR Principles are applicable not only to data objects, but to non-data objects as well, without requiring that data itself be open.[44] This implies that the 5-Star system is slightly more restrictive, at least in terms of what it can be applied against, as it requires that data be fully accessible and free, whereas FAIR merely

[39] Tim Berners-Lee, 'Linked Data: Design Issues', *w3.org*, June 2009, n.pag. <https://www.w3.org/DesignIssues/LinkedData.html> (2 November 2019).

[40] A. Hasnain and D. Rebholz-Schuhmann, 'Assessing FAIR Data Principles Against the 5-Star Open Data Principles', *The Semantic Web: ESWC 2018 Satellite Events*, edited by A. Gangemi et al., Springer International Publishing, 2018, pp. 469–77.

[41] T. Berners-Lee, 'Linked Data: Design Issues', n.pag.

[42] Ibid.

[43] Hasnain and D. Rebholz-Schuhmann, 'Assessing FAIR Data Principles Against the 5-Star Open Data Principles', p. 475.

[44] Ibid., p. 474

requires that in the case of closed or restricted data, the license agreement should be made available to the data consumer through metadata.[45]

The authors conclude that although Linked Open Data may be something to strive for in the future, the current digital research landscape of restricted access, privacy laws, and copyright does not lend itself well to a fully openly accessible online environment.[46] Hasnain and Rebholz-Schuhmann acknowledge that in this sense the FAIR Principles offer a 'broader scope' by including closed data and steps for accessing licensed data, but they do emphasize that true accessibility, a goal for both LOD and FAIR, is hindered by any restrictions on data as a general rule.[47] The authors conclude by labeling the LOD 5-Star system as 'idealistic'; on the other hand, the authors claim that the FAIR Principles have simply reused the 5-Star scheme without reference and merely added in the steps requiring metadata on licensing agreements.[48] Hasnain and Rebholz-Schuhmann do not seem to make a strong claim as to which scheme should be the ideal choice for adoption, citing that either the FAIR Principles will result in reusability through metadata and licensing agreements, or that the 5-Star system will bring us to ultimate reusability because all data should be open and free to begin with.[49] The authors leave it up to the research community to decide for themselves as 'the future will tell,' thus implying that more work needs to be done in either camp when it comes to publishing and reusing scholarly research data.[50]

[45] Ibid.

[46] Ibid., p. 475

[47] Ibid., pp. 475-476.

[48] Ibid., p. 476.

[49] Ibid.

[50] Ibid.

Chapter 2: Research in The Humanities

2.1

Introduction

In the previous chapter we explored the ways in which the FAIR Principles and the Open Science movement have stimulated the conversation about and the adoption of transparency in the scholarly research community. Implementing the FAIR Principles requires not only transparency, but also collaboration and promotion within the community in support of adopting a set of formalized, data-intensive methods and digital tools for best practices. By embracing such standards, it is the hope that Open Science can flourish and aid in the promotion of a fully collaborative, communicative and supportive environment for open data and scholarly research at all levels. However, whether or not these standards can be implemented into humanities scholarship is another question altogether.

In this chapter, we will consider the implications of adopting the FAIR Principles within the Humanities by taking a closer look at the ways in which research is conducted and disseminated amongst its scholars. Any implementation of formal standards within humanities scholarship is complicated by a historical aversion amongst many of its scholars towards collaboration, a reverence for the authority of the monograph, and the inherent challenges in agreeing upon a set of shared terminology or protocols in data collection and dissemination. The consequences of the above is that humanities scholars have been more resistant towards changes than in the natural or life sciences. The multidisciplinary nature of the Humanities in both method and form defines, but also hinders the adoption of standards which could very well transform the discipline; how this could impact its scholarship will be explored here.

2.2

Scholarship within the Humanities

It follows that in order to create a set of standards or principles upon which to base methods and best practices, a clear definition of the subject or object is necessary. Without a strong understanding of the thing we wish to define, there is almost no way in which we can hope to engage critically with it, let alone suggest ways in which to improve or standardize it. Attempting to define the Humanities as a discipline is a challenge much too large to be considered here, however we will explore a set of sufficient definitions as to allow us to standardize its methods within the context of book history.

Christine L. Borgman, Professor of Information Studies at UCLA, argues that 'any lumping of disciplines or domains as "the humanities" is problematic,' not only due to the variety of objects studied by each discipline but also because of the various research methods with which these objects are studied depending on the discipline in question.[51] Everything from the Classics, the Arts, Linguistics, Philosophy, Languages, History, Literature, and even Archeology may be found under the domain of the Humanities, yet all have vastly different approaches to academic research, and the various reasons for how and why these fields are so often placed under the humanities remains at the discretion of those in positions of authority within academic institutions, or else the personal preference of the humanities scholars themselves.[52] For example, in some cases Archeology is categorized as a branch of the Social Sciences instead of the Humanities, and likewise the Arts may either fall under the Humanities, or else under their own domain and joined by fields such as Theatre, Architecture, and Design studies.[53] Furthering the complication, it may be the case that scholars employed in a

[51] C.L. Borgman, *Big Data, Little Data, No Data* (Cambridge, Massachusetts: The MIT Press, 2015), p. 161.

[52] Ibid., p. 200.

[53] Ibid., p. 161.

traditionally non-humanities field or faculty may in fact hold advanced degrees from a humanities discipline, and despite their employment they may still 'self-identify' as a humanities scholar.[54]

The above examples help to illustrate just how varied, dynamic, and fluid the Humanities are as a discipline. As a consequence, humanities scholarship can be equally dynamic and fluid, resulting in the exchange and flow of ideas, methods, and theories passing between disciplines, at times both influencing and enhancing new research methods.[55] Yet, as Borgman writes, despite the potential for the healthy exchange of ideas and research within the Humanities, in general 'humanistic scholarship tends to be more individualistic than in other disciplines.'[56] In complete contrast to the Sciences and the Social Sciences, in which collaboration and 'collective cognition' is both expected and praised, as a discipline the Humanities engage in the least number of instances of 'co-authorship and collaboration' in research.[57] Traditionally, scholarship in the Humanities is introspective, driven by subjective and personal reflection using qualitative analysis, though increasingly there are exceptions, resulting in humanists 'borrowing technologies and methods' from the Social Sciences to conduct quantitative research.[58] Nevertheless, there is real truth behind the image of the 'lone Humanities scholar' quietly pouring over their own research that eventually becomes, usually over a substantial period of time, their nth monograph, and up until recently the notion of collaboration within the Humanities has been both under-discussed and rarely practiced unless a specific project demands so.[59]

[54] Ibid.
[55] Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet* (Cambridge, Massachusetts: The MIT Press, 2007), pp. 212–213.
[56] Ibid., p. 219.
[57] Ibid., p. 219–220.
[58] Borgman, *Big Data, Little Data, No Data*, pp. 161–162.
[59] G. Griffin and M.S. Hayler, 'Collaboration in Digital Humanities Research—Persisting Silences', *Digital Humanities Quarterly*, vol.12, no.1, (2018) <http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html> (1 January 2020).

Humanities scholars typically conduct individual research in the form of close reading and interpretive analysis of specific objects, texts, or ideas, often placing and analyzing them within the context of the cultural periods from which they come to reveal new interpretations and perceptions on society.[60] One scholarly practice that is de rigueur in humanities scholarship is close reading, which can be generally defined as 'a prolonged course of study' involving the concentrated 'scrutiny of a single text or passage' [61] or entire body of work, upon which the scholar employs their interpretive and analytical skills, typically in the field of Literary Studies. The written narrative born from this close analysis often takes the form of a scholarly monograph completed by an individual scholar, an isolated process which continues to be the 'gold standard' in terms of authority and knowledge dissemination within academic circles and publishing in the Humanities.[62] Further, a relatively limited number of scholarly monographs have been fully digitized and placed online, and overall humanities scholars publish less of their work in academic journals or on online platforms, though the balance is shifting as increasingly more Humanities digitization projects are funded.[63] Nonetheless, the above is in direct contrast to the Sciences and Social Sciences, which rely almost exclusively upon collaboration and teamwork to produce their research, as well as publishing the greatest number of research papers and articles in digital scientific and scholarly journals online as a means of spreading knowledge, trust, and authority over a particular area of research.[64]

[60] Borgman, *Scholarship in the Digital Age*, pp. 213–216.
[61] M. Hancher, 'Re: Search and Close Reading', in M. K. Gold & L. F. Klein (eds.), *Debates in the Digital Humanities: 2016* (Minneapolis: University of Minnesota Press, 2016), pg. 122.
[62] Borgman, *Scholarship in the Digital Age*, p. 214
[63] Ibid.
[64] Ibid., pp. 214–215.

2.3

The Humanities Gone Digital

One of the relatively newer fields to emerge within the Humanities is Digital Humanities, a collaborative, multidisciplinary field which has inspired ongoing debates over the years between traditional humanists and their technically-inclined colleagues over everything from semantics, research methods, and authority control, to just what exactly it is that digital humanists do.[65] Definitions of Digital Humanities abound due to the fact that it encompasses a wide range of disciplines and research methods, including Information Technology, Digital and Computational Programming, Social Sciences, Linguistics, Economics, History, Literary Studies, among others; as such, much like the traditional Humanities, Digital Humanities is notoriously difficult to define, and yet many have tried to and continue to attempt to do so. Scholars Lauren F. Klein and Matthew K. Gold loosely define the field as one 'that operates through relation, one that informs and is informed by allied disciplines' and which 'owes its existence to more than one source.'[66] Cambridge University ambitiously defines Digital Humanities as

> a broad field of research and scholarly activity covering not only the use of digital methods by arts and humanities researchers and collaboration by Digital Humanities specialists with computing and scientific disciplines, but also the way in which the arts and humanities offer distinctive insights into the major social and cultural issues raised by the development of digital technologies.[67]

[65] T.E. Clement, 'Where is Methodology in Digital Humanities?', in M. K. Gold & L. F. Klein (eds.), *Debates in the Digital Humanities: 2016* (Minneapolis: University of Minnesota Press, 2016), pp. 153–157.

[66] L. F. Klein and M. Gold, 'Introduction', in M. K. Gold & L. F. Klein (eds.), *Debates in the Digital Humanities: 2016* (Minneapolis: University of Minnesota Press, 2016), p. xi–xii.

[67] 'Defining Digital Humanities', Cambridge Digital Humanities, *cdh.cam.ac.uk*, n.pag. <https://www.cdh.cam.ac.uk/cdh/what-is-dh> (27 January 2020).

Similarly, at the University of Toronto, Woodsworth College, Digital Humanities 'studies human culture – art, literature, history, geography, religion – through computational tools and methodologies,' adding that 'in turn, DH studies the digital through humanist lenses.'[68] The Digital Humanities department at Oxford University includes the 'activities [which] take place all across the Humanities Division, IT Services, The Oxford e-Research Centre, the Oxford Internet Institute, the Bodleian Libraries,' as well as their museums and colleges as a part of their own definition of the field.[69]

Despite variations in language the above definitions lead us to conclude that in general, Digital Humanities aims to combine the affordances of the Humanities with those of Computational Programming and the Sciences in order to better understand and study the world in which we live. The complexity, size, and sheer scope of Digital Humanities as a field does echo similar difficulties that the traditional Humanities continue to face, in that there are scholars and researchers from various educational backgrounds – and used to working under very different conditions and methods – coming together as colleagues to perform and produce research. However, it has not been an easy ride for digital humanists, who have until now 'faced the difficulty of making their work legible to colleagues in their home disciplines' while also dealing with criticism not only from external disciplines who are reluctant to legitimize the field but from within the DH community as well.[70] Several scholars have criticized that the field of DH favours those with 'scholarly status, institutional support, and financial resources,' highlighting just some of the issues surrounding the theoretical inclusivity of the "big tent" of DH.[71] Although there is obviously still much to debate, the field itself has continued to

[68] 'Digital Humanities', University of Toronto, Woodsworth College, *wdw.utoronto.ca*, n.pag. <https://wdw.utoronto.ca/digital-humanities> (28 January 2020).
[69] 'Divisions and Units', Digital Humanities at Oxford, *digital.humanities.ox.ac.uk*, n.pag. <https://digital.humanities.ox.ac.uk/divisions-and-units> (27 January 2020).
[70] Klein and Gold, 'Introduction', p. xi.
[71] Ibid., p. x.

expand and evolve to meet demands for more large-scale, textual and computational analysis methods and tools.[72]

2.4

Data and Collaboration in the (Digital) Humanities

Unlike the Social Sciences, which have 'long studied — and often directly impacted — scholarly information system development' as well as scholarly cyberinfrastructure in the Sciences by adopting scientific methods and digital tools for analysis and research, humanists are still finding their footing when it comes to adopting similar methods, let alone 'expressing how these forms of study map to [their] theoretical concerns.'[73] Scholar Tanya E. Clement claims that in order to ease the identity crisis of DH, conversations surrounding '[h]ow we validate and share knowledge between epistemological frameworks,' whether it be the Humanities or the Social Sciences, 'has much to do with how we articulate the link between our methods and our theories.'[74] With the recent rise in the promotion and support (financial and otherwise) of collaborative, multi-disciplinary projects dealing with "big data" from both governments and institutions alike, teams of researchers and scholars from various disciplinary backgrounds have had to put familiar methods aside and learn not only how to collaborate and communicate amongst themselves, but to do so while navigating an atmosphere of 'otherness' and unfamiliarity in using novel methods as well.[75]

The researchers Gabriele Griffin and Matt Steven surveyed various publications from within the field of Digital Humanities in order to analyze the methods of collaboration currently in use as well as the ways in which collaboration remains largely 'under-developed in both theory and practice,' focusing on the partnerships between academic and non-academic collaborators (technicians) and

[72] Ibid., p. x–xii.
[73] Clement, 'Where is Methodology in Digital Humanities?', p. 153.
[74] Ibid.
[75] Griffin and Hayler, 'Collaboration in Digital Humanities Research—Persisting Silences', n.pag.

their digital research tools.[76] They identify three general types of interactions through which collaboration and information may travel: human-human interactions, human-machine/material interactions, and machine/material-machine/material interactions.[77] Griffin and Steven go on to discuss the most complicated of these relationships, that is the "human-human interactions" between digital and/or computer programmers and humanities scholars, stating that feelings of alterity, destabilization of power relations, and defamiliarization of their own work creates unnecessary tensions between the collaborators.[78] They state that the team 'needs to be built from first principles, emphasizing the skills that brought the collaborators together whilst minimizing the friction of competing disciplinary norms' to create a new common ground from which collaborating disciplines can produce meaningful and novel research.[79] However, Griffin and Steven admit that humanities scholars are not only slow to adopt the use of digital tools and technologies, but also to work as part of a team, which is due in part to the enduring 'Humanities tradition that locates agency, originality, and meaning-making firmly with the author/maker,' further placing their work outside of the realm of the collaborative.[80]

2.5

Lack of Standardization

This noticeable resistance within the Humanities towards digital publications and collaboration has created issues in terms of peer review, quality control, and authority within the discipline, at times perpetuating how humanities scholars have reacted to the shift in recent years from publishing the majority of their work as physical books versus new external pressures to work in teams and publish their research online, changing lengthier formats into shorter versions for articles or even

[76] Ibid.
[77] Ibid.
[78] Ibid.
[79] Ibid.
[80] Ibid.

blog posts. Examples of external pressures include current ‘changes in intellectual property policy, [and] the economics of scholarly publishing,’ as well as the increasing involvement of stakeholders in the form of governments and funders that are escalating tensions ‘between ease of access and the desire to control that access’ of both research data and funding connected to projects, all of which has resulted in a complete restructuring of the sociotechnical model for academic publishing.[81] However, this shift in the organization of academic publishing is seen as both inevitable and necessary if it is to endure into the digital age. It is undeniable that the ways in which humanities scholars now communicate both publicly and privately has changed dramatically because of digital technology. As Borgman states, however, the ‘underlying processes and functions of communication have changed little,’ and in order to survive it may be necessary to work with these external factors rather than against them.[82]

It is interesting to note that humanities scholars 'draw on the longest literature time span of any discipline,' but due to the fact that almost the entire span of humanities scholarship pre-dates the Internet, they actually have 'the smallest proportion of their literature online of any discipline.'[83] Nonetheless, similar to most scholars and researchers today, humanities scholars do of course take advantage of digital technologies and digitized information to perform and compose their research, since it is near impossible now to survive in academia without a computer and an internet connection, regardless of one’s preference for the analog. Borgman points out that due to the reorganization of scholarly book publishing in particular towards digital platforms in recent years, scholarly communication within the Humanities is similarly going through a phase of restructuring that has made publishing physical books more challenging, resulting

[81] Borgman, *Scholarship in the Digital Age*, pp. 75–76.
[82] Ibid., pp. 74–75.
[83] Ibid., pp. 214–217.

in a growing number of humanities scholars becoming less opposed to writing journal articles and having their work published online in a digital format.[84]

However, contributing to humanities scholars'—as well as the public's in general—suspicion of digital publications is the fact that now "anyone" is able to publish their work online, bypassing strict protocols of traditional forms of quality control and peer review associated with scholarly communication, which up until now have dominated authority in academic research.[85] Yet recently, issues with the peer review process, such as lack of confidentiality and objectivity, high costs and lengthy approval timelines, as well as the publishing of fraudulent research data, have resulted in divided opinions within the academic publishing community surrounding long-standing perceived notions of authority and validity.[86] This unease, coupled with the rapidly evolving landscape of digital media, has lead scholars to search for new ways in which to 'establish authority and validate scholarly work' through novel methodological perspectives by combining both humanities and social sciences approaches to research in the hopes of creating new techniques and guidelines for best practices.[87]

Although a generous amount of humanist research data and information has been digitized and is available online, there remains a lingering preference for the physical over the digital despite achievements in digitization and democratization of knowledge in recent years.[88] One reason for this is that it is 'difficult to set boundaries on what are and are not potential sources of data for humanities scholarship,' and when it comes to what Borgman calls the 'three "memory institutions"' (museums, libraries, and archives), it often happens that 'each institution will represent and arrange its objects according to its mission.'[89] Currently, libraries are the most standardized in terms of cataloging and presenting

[84] Ibid., p. 214.
[85] Ibid., pp. 47–48.
[86] Ibid., pp., 60–61.
[87] Clement, 'Where is Methodology in Digital Humanities?', p. 160.
[88] Ibid., pp. 166–168.
[89] Ibid., p. 166.

information regarding their textual collections, a task made slightly easier since most of their objects are published books or journals. Archives and museums by contrast are still far from reaching a unified format for representing their collections online due in part to the variety and uniqueness of their objects.[90] This makes research more difficult and time–consuming on the part of the scholar, since an institution may represent and record an object in one way, yet the same object may be represented in a completely different format or context at another institution. This unnecessarily complicates the research process and potentially influences whether or not a particular object or data source is actually useful for scholarship, with everything from 'its form, genre, origin, or degree of transformation from its original state'[91] influencing the humanist scholar on its veracity and usability. The 'balance between authors, publishers, and librarians has shifted radically,' including their publications, in that when digital they are no longer "stable" in (physical) form but are instead 'malleable, mutable, and mobile' and can be distributed and disseminated in a myriad of ways and with varying degrees of standards by anyone who has access to them.[92]

A lack of standardization in the representation of digital objects, including the perceived lack of standardization in terms of peer review and quality control of online scholarly publications, are two major concerns for humanities scholars and highlight the ongoing debate between those in favour of "going digital" and those who prefer a traditionally analog approach. Although the sheer range of unique objects and instances which can be studied and interpreted through a humanistic lens is seemingly endless, there are a variety of reasons why the original object may no longer be available or else cannot be analyzed in person, so that the only choice left to the scholar is to study a digitized version of the original.[93] Therefore, digital collections held in libraries, universities, cultural heritage institutions and archives

[90] Ibid., pp. 166–167.
[91] Ibid., p. 167.
[92] Borgman, *Scholarship in the Digital Age*, p. 48.
[93] Borgman, *Big Data, Little Data, No Data*, pp. 162–163.

are all hugely important to the humanist scholar, as now more than ever scholarship and research can be completed at a scale never before seen within the discipline. Yet, despite the fact that large digital collections can hold up to millions of recorded items and are continuing to expand, humanities projects utilizing these vast stores remain relatively small in comparison, usually involving only one or at most a handful of researchers closely studying specific subcategories of records or objects over the span of months or even years.[94] So often it is true that for the Humanities, 'data sources can be big in volume, very big in variety, but usually are small in terms of velocity.'[95]

2.6

(Book) History is Consistently Inconsistent

The Humanities are currently experiencing as much of a 'data deluge'[96] as other disciplines due to the mass digitization projects of collections in libraries and cultural heritage institutions. Increasingly, digital humanities projects relying on the use and re-use of large amounts of quantitative data, especially in the fields of Literary Studies, History and Linguistics, are being realized. For example, Alexis Weedon, professor of Publishing at University of Bedfordshire, describes how adopting such social sciences methods as statistical and multivariate analysis can be useful as a means of study in subjects such as book history.[97]

Weedon explains that historically, text production was 'frequently aimed at multiplying and spreading its product as much as possible' which also meant that it was vulnerable to 'markets and market forces.'[98] One consequence of this is that historical records of the book trade exist in multitudes in the form of detailed lists of quantities, whether they be fees paid to an author, the cost of supplies like paper or

[94] Ibid., p. 165.
[95] Ibid.
[96] Ibid., p. 161.
[97] A. Weedon, 'The Uses of Quantification,' in S. Eliot and J. Rose (eds.), *A Companion to the History of the Book* (Malden, MA: Wiley-Blackwell Publishing, 2009), pp. 33–49.
[98] Ibid., p. 33.

ink, print runs, sales figures, rates for binding, advertising and distributing, among other quantifiable figures.[99] Weedon states that the quantitative analysis of this historical data has contributed to useful and relevant observations on a variety of subjects, including the 'origin of the codex, the decline of the English Stock, and the distribution of books in eighteenth-century North America' as well.[100]

Simon Eliot, professor of the History of the Book at University of London and co-editor of *A Companion to the History of the Book*, agrees that quantitative analysis can lead to new insights in the study of book history.[101] For Eliot, looking at specific "case studies" is important, for example 'particular titles, demanding authors, ingenious publishers, depressive booksellers, and perverse readers,' but individual case studies alone are not sufficient to be sure 'that you had assembled a reliable sample that did justice at large to the particular period or area that you were studying.'[102] Further, individual case studies require context in the form of the "bigger picture" if they are to convey their own meaning and significance within book history.[103] However, Eliot cautions that the main reason behind applying statistical methods in book history 'must be its usefulness,' which depends 'not just on what the statistics tell us,' but importantly 'how much we can rely on them' as well.[104]

There are real issues to contend with when using historical records as data which can make research in any historical subject difficult. One reason for this is that the motives behind why the information was captured in the first place can affect how it was collected, as the methods used vary widely depending on its original or intended use at the time.[105] This is certainly the case with book historical data in that members from all sectors of the book and book-affiliated

[99] Ibid.
[100] Ibid.
[101] S. Eliot, 'Very Necessary but Not Quite Sufficient: A Personal View of Quantitative Analysis in Book History,' *Book History*, vol. 5, 2002, pp. 283–293.
[102] Ibid. 284.
[103] Ibid.
[104] Ibid., p. 287.
[105] Weedon, 'The Uses of Quantification,' pp.33, 38.

trades including librarians, scholars, and bibliophiles, as well as those who governed and legislated for laws on taxation and copyright, collected information for a variety of reasons and by using vastly different approaches.[106] Everything from accounting records tracking income received and expenses, to lists of books and bibliographies, to legal and administrative correspondence between clients and publishers, represent just a few of the great number of examples of existing book historical data.[107] The most obvious consequence when it comes to the variation in types of data in book history is that it becomes difficult to compare multiple resources with one another, whether because they are completely different objects from the start or else because the data was collected or recorded in a manner that fit the personal preference of the researcher at the time and not for the benefit of future users. This connects to the lack of standardization in the Humanities that has been discussed earlier and is a significant issue in terms of data in general within this discipline.

Although consistency, or the lack thereof, continues to be one of the foremost concerns for all historians and represents the majority of issues when dealing with historical records of any type, undoubtedly it is the quality of the data that is the 'chief problem for book historians wanting to use quantitative methods,' and for a variety of reasons.[108] Weedon explains that the sample size is usually quite small, 'selected for preservation or significance rather than at random,' and that the 'information on how the data were compiled or what they measured is sometimes lost' or at the very least incomplete.[109] She also cites issues such as 'primitive administration, book-keeping, and reporting procedures' as contributing factors to the 'pseudo-statistics in the historical record,' increasing the difficulty for book historians to trust the consistency and reliability of historical data.[110]

[106] Ibid.
[107] Ibid.
[108] Ibid., p. 39.
[109] Ibid.
[110] Ibid.

Likewise, Eliot points out that 'the past has left us some data, but they were not produced in laboratory conditions; they were not designed to answer our questions,' nor are they a representative sample, and they almost never would have used a 'classification system that we might find at all helpful.'[111] In other words, when it comes to historical data, we can forget standards, complete and intact information, reliable terminology, and unified data formats. Historical data are messy, but as Eliot says 'they are all we have got and we must work with them.'[112]

At Erfgoed Leiden en Omstreken, an archive and cultural heritage institution located in Leiden, The Netherlands, it is similarly the quality of the historical data that is the most important, and yet most difficult, to manage.[113] Historical information and objects abound within Erfgoed's archives, but due to a lack of standards in how information with first collected and preserved all those years ago, quite often those involved in transforming analog information into digital are working off of traditional methods such as index cards and handwritten records, all of which are in varying in degrees of completeness, consistency, and reliability.[114]

Depending on the item, it may be labeled as "item n. 1" or "1810" as in the year, or it may simply lack a title at all, and so with no other material to go on the challenge then lies in filling in the rest of the information themselves in a way that accurately describes the item so that they may then catalogue and transform it into digital data that is both useable and useful to publish online.[115] The task of transforming historical data from traditional, analog methods into digital data is hugely time-consuming and difficult, not to mention possible issues concerning copyrights which can also greatly slow down the publishing process.[116] In this way it is also a challenge to complete the metadata as well for the same reasons as

[111] Eliot, 'Very Necessary but Not Quite Sufficient: A Personal View of Quantitative Analysis in Book History,' p. 287.
[112] Ibid.
[113] E. Gehring and W. Hasselo, interview conducted by author, Leiden, 12 December 2019.
[114] Ibid.
[115] Ibid.
[116] Ibid.

mentioned above, sometimes taking years to completely rectify, and so the quality (i.e. the completeness) of the data to begin with is of utmost importance.[117]

Although Erfgoed relies on their own system for transforming and publishing historical data onto their online digital archive, they do take inspiration from both the FAIR Principles and the Linked (Open) Data principles as well.[118] Without specifically claiming one method, they aim to combine the philosophies of each method in what may be termed as 'LOUD', or 'Linked Open Useable Data,' focusing the majority of their energy on ensuring that the data that is published online is not only open and freely accessible to the public, but is also useable and reusable as well.[119] Ultimately, Erfgoed aspires to be able to share all of its historical information online not only with the general public, but with other national, historical and cultural institutions as well, so that scientists, researchers, and scholars alike can access their data and datasets, use them, transform them, and publish their own data online, linking back again to Erfgoed in a completely interoperable and circular framework.[120]

Throughout this chapter we have explored the factors that complicate data sharing in the Humanities. The nature of humanities scholarship itself greatly influences the sharing and dissemination of data, in that humanities scholars are accustomed to performing solitary research and producing monographs, in contrast to participating in team projects and producing co-authored studies as is seen regularly in the Sciences and Social Sciences. Humanities data is also notoriously difficult to define, since it is incredibly diverse and can range wildly not only in form but also in meaning and context depending on its origin and representation within society. These factors will be explored in the following chapter in which we will look

[117] Ibid.
[118] Ibid.
[119] Ibid., 'LOUD' as termed by W. Hasselo.
[120] Ibid.

at a specific case study and attempt to FAIRify a data resource according to the FAIRification Process.

Chapter 3: Case Study

3.1

Introduction and Description of the Case Study

The aim with this case study is to illustrate the process of "FAIRifying" a database from within the field of book history. It will be a critical investigation into the practical limitations and challenges that arise when applying the FAIR Principles in this case, onto a database from the Bibliotheca Thysiana, a seventeenth-century library located in Leiden, The Netherlands. The database is the result of information collected by Esther Mourits[121] while conducting research during her PhD, in which she reconstructed how the library was assembled and in turn investigated the reasons why its founder, the lawyer Johannes Thysius (1622-1653), left his substantial book collection to the public upon his death.[122]

Upon agreeing that her database could be used for this case study, Mourits also expressed a desire to see the information from her research published so that it may become useful for future scholars.[123] Mourits explains in an email that the database was born out of practical purposes to answer her research questions, built up by using information she collected from the oldest catalogue of the Thysiana as well as handwritten notes left by Thysius in his cashbook.[124] Mourits admits that the database is incomplete in its current state, as she was not always successful in identifying books noted by Thysius as existing titles, and so in those instances she left all fields in the database other than the title empty.[125] As well, a significant portion of books purchased were never actually held in the Thysiana but are still included in the database because they were noted as having been purchased by

[121] See the commercial publication of Esther Mourits: *Een Kamer Gevuld Met De Mooiste Boeken: De Bibliotheek Van Johannes Thysius (1622-1653)*, (Uitgeverij Vantilt, 2016).

[122] P. Hoftijzer, 'Dissertation on the life and library of Johannes Thysius', universiteitleiden.nl, 14 December 2016, <https://www.universiteitleiden.nl/en/news/2016/12/dissertation-on-the-life-and-book-collection-of-johannes-thysius> (8 February 2020).

[123] Email correspondence with Esther Mourits. (2 October 2019).

[124] Ibid.

[125] Ibid.

Thysius himself; on the other hand, books that were purchased and included in the Thysiana after his death are not included in the database.[126] In this way, the database is not a catalogue of the current contents of the Thysiana but a collection of data depicting the books directly purchased by Thysius. Further, not all information is completely accurate due to the fact that it is based on archival material versus the library itself, which thus led to some instances of interpretation at the time of data collection.[127] As Mourits explains, due to constraints on access and time to physically visit the Thysiana and describe each book, title information for a great number of books was taken directly from the library catalogue of Leiden University, and if the books were no longer in the library she searched through international catalogues to find the title information.[128]

3.2

Brief Description of the Database

The database, which was created using the program FileMaker Pro 12, contains 3479 files and is written in Dutch. The information is divided into nine categories using the following terms: ‘oud volgnummer’ (old serial number); ‘auteur’ (author); ‘titel’ (title); ‘plaats van uitgave’ (place of publication); ‘jaar van uitgave’ (year of publication); ‘uitgever’ (publisher); ‘formaat’ (format of the book, ex. folio, quarto, octavo, etc.); ‘prijs in guldens en stuivers’ (price in guilders and pennies); ‘boekbinder’ (bookbinder). Although the language of the database is written in Dutch, it should be noted that the language of such information as titles, place names and publishers are represented in their original language, ranging from Latin, German, French, or Dutch depending on the book in question.

Upon investigation into the quality of the data, a number of important issues arise that are worth noting before we attempt to FAIRify the data. One of the major issues that is immediately apparent are the numerous empty cells where no

[126] Ibid.
[127] Ibid.
[128] Ibid.

information has been recorded at all, which as Mourits explained is the result of being unable to identify or find the information with the resources she had at the time of data collection. Figure 1 below shows just how much data is missing, for example in the column 'boekbinder,' over 3000 cells are empty, about 86.2% of the total number of books in the database, and in 'uitgever' just over 1000 cells are empty, or 29% of the total number of books. This is significant because missing data can negatively impact future data analysis by not showing an accurate representation of the objects being studied. If the data is incomplete, then the analysis is by default incomplete as well.

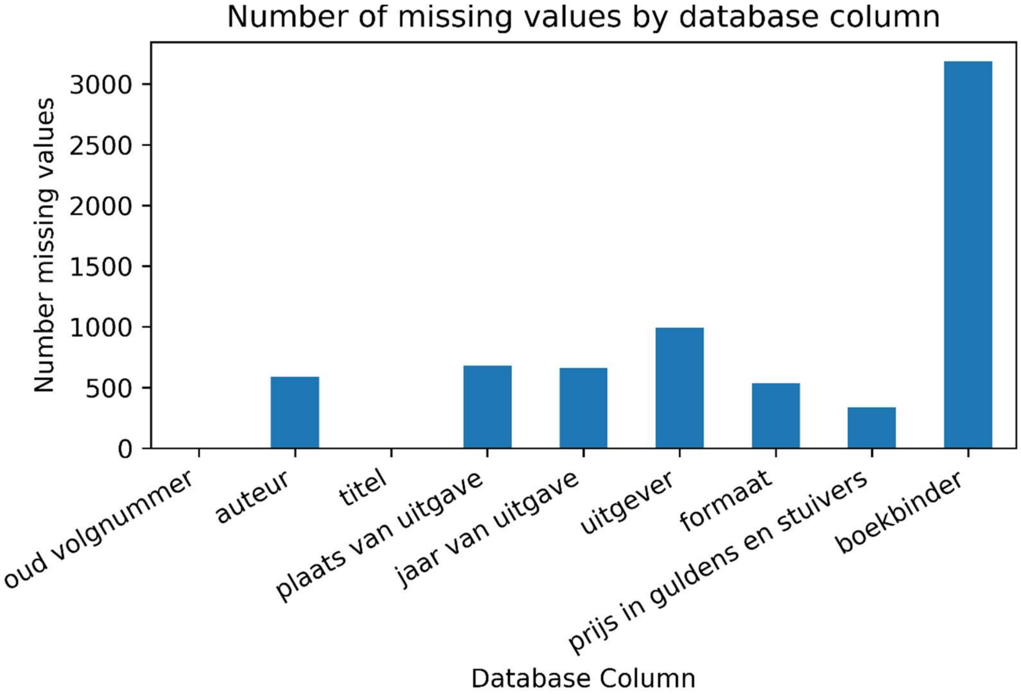

Figure 1: The number of missing values in the Thysiana database per column.

Another issue to note is the inclusion of values such as '0' in cells where the information was either unknown or else entered for some other unknown reason during data collection. Figure 2 demonstrates the number of books published each year, yet as we can see there are a number of books that were apparently published around the year zero, which is certainly unrealistic and necessitates additional clarification, further solidifying that there are data quality issues.

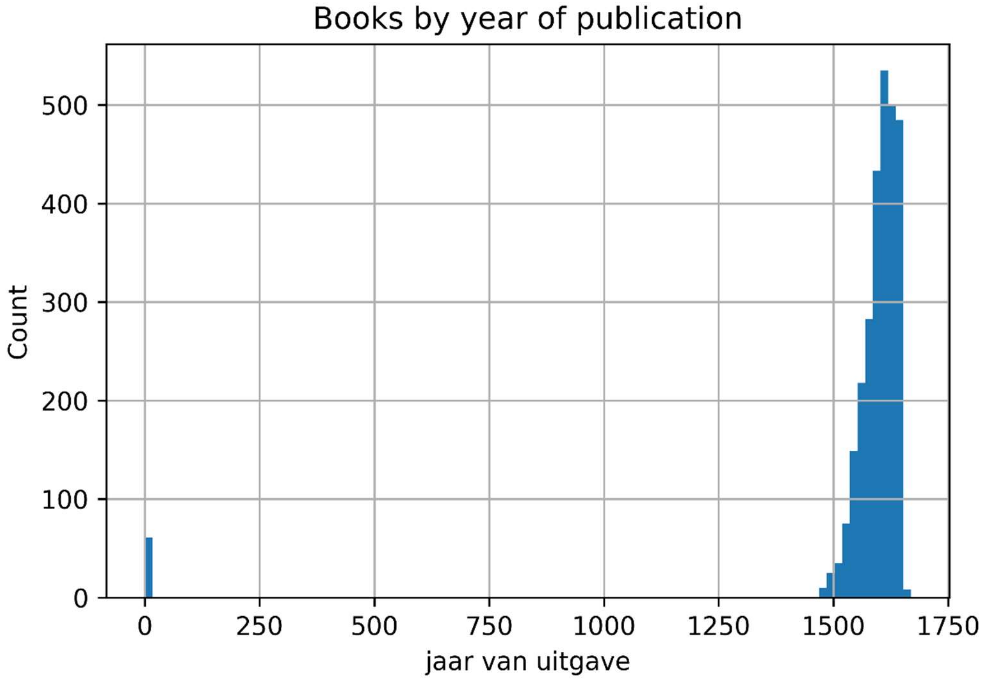

Figure 2: The number of books published by year.

The final example is displayed in Figure 3, in which we can see the total number of books that were published in the top 32 cities of the database. As is evident, we again have data quality issues in that some of the cities are recorded more than once but with square brackets, a simple addition that could be interpreted as a mistake in inconsistency but which may indicate that the information has been supplied from an external source; nonetheless, this which would meaningfully affect the accuracy of results through any sort of data analysis.

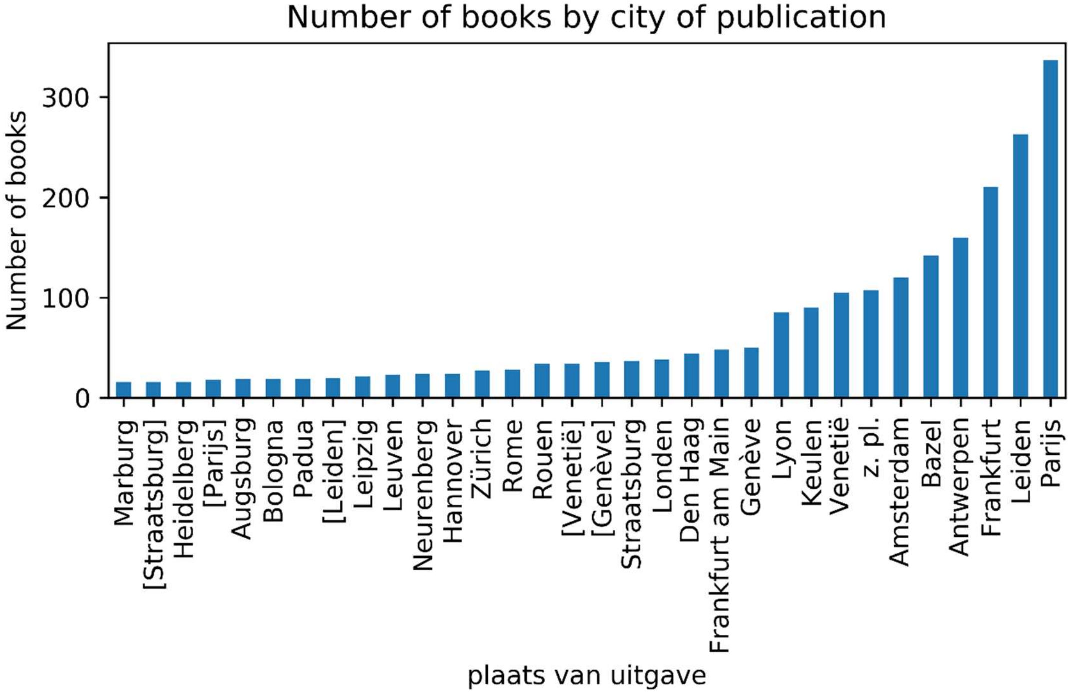

Figure 3: The number of books published by city.

These data quality issues need to be addressed before FAIRifying the data and especially so in Figures 2 and 3 where not only is there missing data but its missing in inconsistent ways or is otherwise repeated. Nonetheless, the historical significance of this database should not be overlooked, as it has the potential to provide future scholars and researchers with rare and valuable data on publishing, private book collections, and book trade in the seventeenth-century, not only in the Netherlands but abroad as well, leading to even more projects and meaningful discoveries in the field of book history. Therefore, FAIRifying the Thysiana database is the first step towards ensuring its reuse and influence upon scholarly research.

3.3

Properties of the Database with Respect to FAIR

Looking back to the FAIR Principles as outlined in Chapter 1, we can now assess the adherence of the Thysiana database to each FAIR recommendation one principle at a time, with a particular focus on the specific properties currently hindering its reuse.

The first principle we will explore is Findable, more specifically F4. The ability to successfully locate data whether through richly described metadata or PIDs is necessary but ultimately not satisfactory enough if the (meta)data itself is not 'registered or indexed in a sustainable resource' or placed within a searchable repository, which when applied to our case study it completely misses the mark. Due to the fact that the database was created for a specific project, the purpose of which was fulfilled once the project ended, the database was kept on a personal computer and was not preserved in such a way that would make it findable by either humans or machines in the long-term, a requirement of principle F4. The reality is that valuable data resources regularly go unused and undiscovered, often sitting on the personal hard drives or USB sticks of well-intentioned researchers, regrettably deteriorating in both functionality and relevance the longer that they are forgotten. This crucial step in the FAIRifying process means that without having directly communicated with the owner of the Thysiana database, it would not have been possible to "find" the database in the first place, obviously hindering its reuse. Further, the database does not contain any documentation or metadata, both of which are sources of information which can aid the human user or machine in locating specific data resources; this requirement will be explored more fully in Interoperability.

The next principle is Accessible, where we take a closer look at A1.1. This principle requires that the protocol through which data is retrieved must be 'open, free, and universally implementable' so that anyone may access the (meta)data through commonly-used platforms that are both free, i.e. cost nothing, and open-sourced. However, the Thysiana database was created using the proprietary software program FileMaker Pro 12 which requires the user to purchase a license in order to utilize it. Similarly, if we look again at Berners-Lee's LOD 5-Star principles, achieving 3 stars requires the use of non-proprietary formats, such as using CSV. Ensuring that the accessibility of data resources is as unrestricted as possible, or at the very least making available the relevant metadata indicating how to recover the data if private or restricted, is an important step in guaranteeing reuse.

We now move onto Interoperability, which is potentially the most challenging of the principles for which to aspire, at least in terms of our case study and the data quality issues explored thus far. Interoperability in FAIR relies upon the use of generic and widely used languages, keywords, and ontologies in open, machine-readable formats to represent knowledge and information, thus guaranteeing the automatic searchability and connectivity of (meta)data. Principle I1 states that (meta)data should be presented in a 'formal, accessible, shared, and broadly applicable language' that is understood not only human to human, but machine to machine as well. As was previously mentioned, our database does not contain documentation or metadata of any type, meaning that "what you see is what you get" in terms of what information has been included and how or why the database has been created. This is the first crucial issue encountered with the Thysiana database, greatly influencing how a machine or human would potentially interact with or understand the nature of the data.

As discussed in section 3.2, one of the most problematic issues is not only the lack of documentation or metadata, but also the inconsistencies in the collection of the data as well, seen in the use of plain text, the recording of unnecessary values or symbols, the long and multi-lingual book titles, as well as variations in the name of the publication city. Although it is not unheard of to use plain, descriptive text when gathering data, by including documentation explaining the methods behind these choices scholars intending to reuse the data are then not forced into interpretation or confusion and as such would greatly increase the reusability and interoperability of the database.

Due to the nature of book publishing in the seventeenth-century, oftentimes the name of the publisher ('uitgever') is represented in a different language depending on which city or in what language the book was published, so that they may be referred to as "Elzeviriana" in a book that was published in Latin, or else as "Elsevier" if the book was published in the city of Leiden, both instances of which we can see in the database. For someone familiar with the nuances of seventeenth-century book-publishing, encountering the same publisher or author name in multiple forms would not be an issue, however without historical context or explanation we run the danger that someone less experienced in the field may interpret these publisher names as two separate entities instead of one in their own analysis of the database.

Another concern of interoperability from the 'uitgever' or publisher field is evidenced by the inclusion of more than one publisher or printer together in one cell, for example "Jean Berthelin, à l'imprimerie de Jean Durand," done so because they had in fact worked together on one book, or else the name of the publisher may be abbreviated in one cell as "Henr. Petri" but is displayed in the next as "[Henr. Petri]" with square brackets. Similarly, as was already pointed out in the previous section place names are irregularly recorded in that they are written as both "Venetië" and "[Venetië]" with two square brackets, and represented as the Dutch translation of the city name instead of the English version. The use of one's native language is not overtly discouraged, however for purposes of interoperability the complication lies in how difficult it becomes to automatically search for commonly used terms, keywords, and names from which to conduct searches. Transforming these differences in language into the more common scholarly English would enhance its use and reuse in a variety of environments, especially in terms of creating linked data, as we will see in the next section. In this situation, it would be useful to record two versions of the place name, one that captures the original form that was used in the original historical document, and the in the form of the standardized reference to the city name.

As we saw in Figure 1 there are many fields that are left completely empty when the information is unknown, but they may also contain a “[?]” or other denotation such as “0,” which likewise complicates the automatic searchability of the database through the increased potential to bring up incorrect variables and results. As well, in some instances, such as in the ‘boekbinder’ field, there are complete sentences including symbols and numbers without further explanation or justification of their use, for example: ‘Eén deel is door Wolter de Haes 'vermaect' voor fl. - : 8 (1-12-1651, rekening ABT 101 D1).’ This would undoubtedly be difficult for the person or machine entrusted to parse through this information, not only because it is recorded in plain text but also because the symbols and values are easily confused or misunderstood. One final obvious inconsistency is found in the ‘prijs in guldens and stuivers’ field for the price of the book, in which the dash symbol is placed in front of some of the values but not all. It may be assumed that because it is included beside a number that it represents a negative value, yet it may also represent the number zero, where if the price begins with a dash then the price was only a particular number of ‘stuivers.’ However, this is certainly speculation and is something else that would need to be clarified with documentation.

The final principle to be explored is Reusable. Achieving the ‘R’ in FAIR ensures that (meta)data are sustainable, accessible, richly described, and findable, all of which leads to data reuse. Importantly, R1.1 focuses on the need for transparency in the usage licenses of data resources, since licensing restrictions can complicate automated searches and render the data useless to those who do not meet specific requirements to access the data. Not only does the database not include mention of usage licenses, but also due to the proprietary programs used to represent the data, this principle has clearly not been met. The same can be said for principle R1.3, in which (meta)data must meet ‘domain-relevant community standards.’ As was previously discussed above, the lack of adherence to a controlled standard, community or otherwise, greatly diminishes the chances that the Thysiana database in its current form can or will be reused by scholars, not without quite a lot of work to bring it up to FAIR standards. Achieving ‘use(ability) for the community’ is a priority of FAIR, which means that even if a researcher chooses to use unconventional or personal methods and tools to capture their data, as Mourits has done with her database, the reasons behind why they chose these unique methods and tools should be clearly stated in either documentation or metadata, neither of which has been done here.

3.4

How Do We FAIRify Our Data?

As stated on the *Go FAIR* website, although the majority of the conditions for findability and accessibility can be met through the quality of the metadata, interoperability and reuse 'require more efforts at the data level.'[129] The process of transforming the unstructured and undocumented Thysiana database into a FAIR data resource can be made easier by following Go FAIR's 'FAIRification Process' scheme, as shown in Figure 4, which will serve as a guide for this next section.

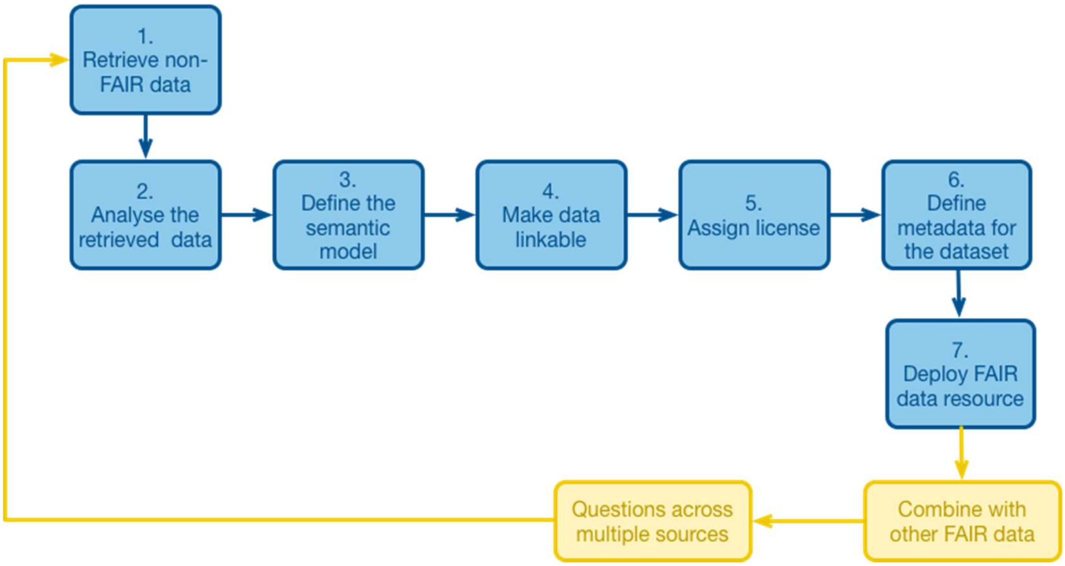
1.
Retrieve non-FAIR data
2.
Analyse the retrieved data
3.
Define the semantic model
4.
Make data linkable
5.
Assign license
6.
Define metadata for the dataset
7.
Deploy FAIR data resource
Combine with other FAIR data
Questions across multiple sources

Figure 4: FAIRification Process. Source: *Go FAIR.org*.[130]

According to the FAIRification Process, the first two steps have already been accomplished through receiving the non-FAIR data and then analyzing the quality and structure of the data as was done in the previous sections. The next step is to 'define the semantic model,' which requires that we 'describe the meaning of entities and relations in the dataset accurately, unambiguously, and in a computer-actionable way.'[131] It is recommended to use existing models so that they are

[129] Anon., 'FAIRification Process', *GO FAIR.org*, n.pag. <https://www.go-fair.org/fair-principles/fairification-process/> (11 February 2020).

[130] Ibid.

[131] Ibid.

representative of a widely-used and agreed upon set of terms, ontologies, and vocabularies as per a specific domain.[132]

As a part of the Top 10 FAIR Data & Software Global Sprint that was held in 2018, the authors of the paper concentrating on 'Humanities: Historical Research' suggest that a good place to start when creating a data model is to 'explore whether some of the general topics that you focus on have already been assigned persistent identifiers or URIs,' and implement them into your own data model in order to 'make it clear that we are talking about the same thing when we exchange knowledge.'[133] They give a few examples of well-known ontologies and shared vocabularies, such as FABIO and the Bibliographic Ontology (BIBO), both of which are used for describing the aspects of books.[134] As for our case study, defining the semantic data model involved first listing the important entities and their relationships to be classified through ontologies and shared vocabularies, for example, 'auteur,' 'titel,' 'boekbinder,' and 'taal' (language), to name just a few.[135]

After identifying the most integral entities and relationships, the next phase in the FAIRification Process is to 'make linkable data.' This is achieved through implementing the data model we have just created with Semantic Web and Linked Data technologies.[136] According to the *GO FAIR* website, this step 'promotes interoperability and reuse, facilitating the integration of the data with other types of data and systems.'[137] We used the following ontologies and shared vocabularies in order to classify and define our entities: FABIO,[138] The Dublin Core Schema

[132] Ibid.

[133] K. Hettne, P. Verhaar, et. al, 'Humanities: Historical Research', in C. Erdmann, N. Simons, R. Otsuji, S. Labou, R. Johnson, G. Castelao, … T. Dennis, *Top 10 FAIR Data & Software Things*, Zenodo., February 2019 <http://doi.org/10.5281/zenodo.2555498> (11 February 2020).

[134] Ibid., p. 40.

[135] For the ongoing FAIRification Process see the Thysiana GitHub account: P. Verhaar, Thysiana, (2020), GitHub repository <https://github.com/peterverhaar/thysiana> (17 February 2020).

[136] Anon., 'FAIRification Process'.

[137] Ibid.

[138] See https://w3id.org/spar/fabio

(dcterms),[139] Schema.org,[140] The Worldcat ontology,[141] BIBFRAME,[142] BIBO,[143] OntoMedia,[144] and FRAPO.[145] Entities and relationships in our data model thus transformed from 'auteur' into 'schema:author,' and 'titel' into 'dcterms:title,' etc., now defined through a machine-readable and actionable term describing their meaning and relation to the other entities in the database that are also semantically easy to understand by the human researcher. For our case study, this step is a work in progress as not all entities have been successfully matched with an appropriate property from existing ontologies. In case it is not possible to find a matching property, which can be the case when dealing with data that is highly specialized or semantic, we would then need to propose a property ourselves and prefix it with an associated namespace of our own making.[146]

The next step in making linkable data is to take the data model and, together with the raw data from our database, convert the data into to the Resource Description Framework (RDF) format, a 'technology which enables you to publish the contents of a database via the web.'[147] This is achieved by recording the data in what are called RDF triples, a data model which 'assumes all statements about resources can be reduced to a basic form consisting of a subject, a predicate, and an object.'[148] In this way, the ontology acts as a sort of "dictionary" from which we take the classified entities and relationships and "tag" them in RDF using linkable data in the form of URIs and PIDs. By storing the data 'in a format in which […] their

[139] See http://purl.org/dc/terms/

[140] See http://schema.org/

[141] See http://experimental.worldcat.org/ontology/library/

[142] See http://id.loc.gov/ontologies/bibframe/

[143] See http://purl.org/ontology/bibo/

[144] See http://contextus.net/ontology/ontomedia

[145] See http://purl.org/cerif/frapo/

[146] See the website Linked Open Vocabularies for examples of existing vocabularies and ontologies: https://lov.linkeddata.es/dataset/lov/

[147] K. Hettne, P. Verhaar, et. al., 'Humanities: Historical Research', p. 40.

[148] Ibid., pp. 40-41.

properties and their characteristics are identified using URIs' as often as possible, it ensures the persistency, reuse, and interoperability of the data elements.[149]

After we have defined the semantic data model and created linkable data, the next phase is to assign a license to the database by referring back to principle R1.1, which states that usage rights should be clearly and explicitly specified and attached to your data.[150] Even though assigning a license is a part of creating rich metadata, it has been specifically included in the FAIRification Process because it is an important step in advertising whether or not a user can access and reuse the data, regardless if it is presented as open access.[151] Likewise, satisfying the last two steps in the FAIRification Process relies upon following the advice outlined in the FAIR principles concerning the creation of rich metadata in order to finally 'define metadata for the dataset' and eventually publish or 'deploy [the] FAIR data resource' with its attached metadata and license.[152] The data resource should be 'indexed by search engines' in order to ensure that the data is remains accessible, 'even if authentication and authorization are required.'[153]

For our case study, due to time constraints not all steps within the FAIRification Process have been fully achieved. However, it is an ongoing process with a goal set to eventually complete each stage in order to finally publish a fully FAIRified database of Mourits' research for the Bibliotheca Thysiana. In light of what we have explored in this section, completely FAIRifying the database will indeed take a concerted effort in terms of ensuring interoperability due to the intensely semantic nature of the database. In the final chapter, we will revisit the implications and challenges encountered while attempting to FAIRify this database, as well as explore some of the ways in which FAIR might be useful to humanities scholarship.

[149] Ibid.
[150] Anon., 'FAIR Principles – Reusable'.
[151] Anon., 'FAIRification Process'.
[152] Ibid.
[153] Ibid.

Chapter 4: FAIRifying for the Future

4.1

Can We Achieve FAIR with Humanities Data?

After exploring the implications and methods of applying the FAIR Principles to a database from the field of book history, we can say that FAIRifying such a database is in fact feasible and worthwhile, though not without challenges. We will now examine some of these challenges below as well as provide recommendations on how to improve conditions in order to achieve FAIR.

Looking back to Chapter 3, our attempt to FAIRify the Thysiana database will continue as an ongoing process. Most significantly, there is still a substantial amount of work needed to bring the quality of the data up to a level where we can apply Semantic Web and Linked Data technologies to eventually convert the data into RDF for publishing on the web. As previously stated, these are key steps in the FAIRification process, relying upon the use of common ontologies, keywords, and shared vocabularies in machine-readable formats in order to be achieved. It is perhaps pertinent to recall as well that according to the FAIR Principles authors, the main reason why data collection takes so long is not because of a lack of available technology, rather it is that we do not treat our 'valuable digital objects' with the kind of attention and care that they require at the time of creation and preservation which renders them difficult to find and ultimately (re)use.[154] The authors are certainly referring to, among others factors, metadata and documentation, and the undeniable emphasis that the principles place on metadata highlights one of the main reasons why it was a challenge to apply each FAIR principle successfully to our database. As a result, Interoperability has been the most difficult principle to achieve during FAIRification.

[154] Wilkinson, et al., 'The FAIR Guiding Principles for scientific data management and stewardship', p. 3.

It would be unduly fatalist to suggest that because we did not fully FAIRify the database that it is not feasible to do so. Instead, perhaps our definition of what a FAIR data resource looks like needs to become more flexible, especially so now that FAIR is being applied to a growing number of digital objects from a variety of disciplines. It would be unrealistic to expect any data resource to reach a point of complete FAIRness, as the data will continue to transform with each (re)use, and as digital technology itself continues to advance and evolve, digital objects will themselves change and evolve as well, necessitating the exact kind of flexibility and open discussion surrounding preserving digital resources that FAIR promotes. After all, the FAIRification Process is referred to as a "process," so it stands to reason that it is actually a continuum, moving between data that are more FAIR versus less FAIR, without a definitive end.

Moving forward, our goal is to improve upon the quality of the database through enriching the existing data as linkable and machine-readable, as well as adding rich, detailed metadata to publish alongside it. It would then be a matter of converting the rest of the data into RDF, assigning a usage license, and publishing the database and the metadata to a persisting repository online. However, this is only if we decide to publish the database in its current and incomplete state, whereas if we attempt to identify all of the missing information the process would take considerably longer. As Mourits stated, she had to utilize international catalogues just to find title information for the books that she could not identify as existing, and so it would be a project in of itself to complete the database in full.[155] Looking again to Erfgoed Leiden, the researchers work with low quality data and missing or incomplete information on a regular basis which dramatically lengthens the amount of time spent on improving the quality of the data for publication, resulting in some digitization projects taking multiple years to complete.[156] With this in mind, it is safe to say that completely transforming the Thysiana database

[155] Mourits, 'Vraag over database Bibliotheca Thysiana', email correspondence.
[156] Gehring and Hasselo, interview conducted by author.

as FAIR could take a similar length of time if we were to seek out and attempt to rectify the missing information.

One possibility that could potentially streamline the FAIRification process would by be asking humanities scholars to plan for the creation and preservation of FAIR data before they even begin generating data resources. By planning ahead and thinking about the steps that need to be taken during the collection and recording of data, including tracking the methods and digital tools used during the process through detailed documentation, the researcher can satisfy many of the FAIR Principles before they have even finished their project, making it easier on themselves and future scholars when they eventually publish the data. It is simply asking that humanities scholars take more responsibility for the digital data that they produce, however this can also have implications when we think about the role that humanities scholars have traditionally occupied.

The question must be asked whether or not we should expect all humanities scholars to become *digital* humanities scholars, or if it is preferable instead to promote the acceptance and utilization of professional digital humanists and technicians to work alongside them through each phase of the data production and publishing process. If humanities scholars remain true to the tradition of conducting research alone, they will need to attain a certain level of digital competence in order to produce and share their research data online, the implications of which can result in poorly executed data production simply due to the fact that they lack the level of experience with and expertise in data production that a dedicated digital technology professional possesses. On the other hand, promoting collaboration in the production of scholarly research data through the inclusion of digital technology professionals from the humanities and other disciplines would almost guarantee higher quality data resources, and with increasing importance placed on funding “big data” projects requiring multi-disciplinary teams and extensive data management plans, the practicalities of dividing tasks to those most capable has its advantages.

The difficulties experienced in our attempt to apply the FAIR Principles to the database reflects similar difficulties that (digital) humanities scholars currently face when endeavoring to define their discipline; first, there are many different versions of the same definition when it comes to what humanities scholars, digital or not, actually do, and second, there are countless ways to describe humanities objects whether they be a book, a painting, or a piece of music, and in this way it is easy to understand the issues that we encountered with our database. Without a clear consensus amongst the scholarly community as to what humanities objects of study should or can be defined as, including how or why we should be thinking about best practices in terms of sharable and reusable research data, it will continue to be a challenge to implement guidelines such as the FAIR Principles to publish scholarly data online.

There is a reason why the FAIR Principles have been widely supported and promoted within the scholarly research community as well as research initiatives such as the EU Framework Program Horizon 2020: the principles offer a flexible yet tangible summary of actionable steps for publishing scholarly research data without forcing the researchers into adopting restrictive standards that may “pigeon-hole” their research into unyielding formats. Their emphasis on machine-actionability prepares scholarly research data for the future in which we will be dealing with larger and more complex data-types, ensuring continuous interoperability and reusability of valuable research data. Digital research in fields such as book history can greatly benefit from the increased interoperability that FAIR has the potential to provide, as there is an unending amount of undiscovered and unused data waiting to be utilized in a variety of projects if only researchers could access and use them. As research in the Humanities continues to evolve towards an increased use of digital technology and platforms, it is essential now more than ever that we create and preserve data in formats that are findable, accessible, interoperable, and ultimately reusable.

Bibliography

Anon., 'Defining Digital Humanities', Cambridge Digital Humanities, *cdh.cam.ac.uk*, n.pag. <https://www.cdh.cam.ac.uk/cdh/what-is-dh> (27 January 2020).

Anon., 'Digital Humanities', University of Toronto, Woodsworth College, *wdw.utoronto.ca*, n.pag. <https://wdw.utoronto.ca/digital-humanities> (28 January 2020).

Anon., 'Divisions and Units', Digital Humanities at Oxford, *digital.humanities.ox.ac.uk*, n.pag. <https://digital.humanities.ox.ac.uk/divisions-and-units> (27 January 2020).

Anon., 'FAIR Principles', *GO FAIR,* n.pag. <https://www.go-fair.org/fair-principles/> (30 October 2019).

Anon., 'FAIRification Process', *GO FAIR.org*, n.pag. <https://www.go-fair.org/fair-principles/fairification-process/> (11 February 2020).

Anon., 'H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020', *European Commission Directorate-General for Research & Innovation*, July 2016, pp.1-12, <https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf> (2 November 2019).

Berners-Lee, T., 'Linked Data: Design Issues', *w3.org*, June 2009, n.pag. <https://www.w3.org/DesignIssues/LinkedData.html> (2 November 2019).

Boeckhout, M., et al, 'The FAIR Guiding Principles for Data Stewardship: Fair Enough?' *European Journal of Human Genetics*, vol. 26, no. 7, July 2018, pp. 931–36. *www.nature.com* <doi:10.1038/s41431-018-0160-0> (1 November 2019).

Borgman, C.L., *Big Data, Little Data, No Data*, (Cambridge, Massachusetts: The MIT Press, 2015).

__., *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, (Cambridge, Massachusetts: The MIT Press, 2007).

Clement, T.E., ‘Where is Methodology in Digital Humanities?’, in M.K. Gold & L.F. Klein (eds.), *Debates in the Digital Humanities: 2016*, (Minneapolis: University of Minnesota Press, 2016), pp. 153–157.

Eliot, S., ‘Very Necessary but Not Quite Sufficient: A Personal View of Quantitative Analysis in Book History,’ *Book History*, vol. 5, 2002, pp. 283–293.

Gehring, E., and W. Hasselo, interview conducted by author, Leiden, 12 December 2019.

Griffin, G., and M.S. Hayler, ‘Collaboration in Digital Humanities Research—Persisting Silences’, *Digital Humanities Quarterly*, vol.12, no.1, 2018, <http://www.digitalhumanities.org/dhq/vol/12/1/000351/000351.html> (1 January 2020).

Hancher, M., ‘Re: Search and Close Reading’, in M. K. Gold & L. F. Klein (eds.), *Debates in the Digital Humanities: 2016*, (Minneapolis: University of Minnesota Press, 2016), pg. 122.

Hasnain, A., and D. Rebholz-Schuhmann, ‘Assessing FAIR Data Principles Against the 5-Star Open Data Principles’, *The Semantic Web: ESWC 2018 Satellite Events*, edited by A. Gangemi et al., Springer International Publishing, 2018, pp. 469–77.

Hettne, K., P. Verhaar, et. al, ‘Humanities: Historical Research’, in C. Erdmann, N. Simons, R. Otsuji, S. Labou, R. Johnson, G. Castelao, … T. Dennis, *Top 10 FAIR Data & Software Things*, Zenodo. February, 2019 <http://doi.org/10.5281/zenodo.2555498> (11 February 2020).

Hoftijzer, P., ‘Dissertation on the life and library of Johannes Thysius’, universiteitleiden.nl, 14 December 2016 <https://www.universiteitleiden.nl/en/news/2016/12/dissertation-on-the-life-and-book-collection-of-johannes-thysius> (8 February 2020).

Klein, L.F., and M. Gold, ‘Introduction’, in M. K. Gold & L.F. Klein (eds.), *Debates in the Digital Humanities: 2016*, (Minneapolis: University of Minnesota Press, 2016), p. xi–xii.

Lahti, L., N. Ilomaki, M. Tolonen, ‘A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800’, *Liber Quarterly*, 2 (2015), pp. 87-116, <10.18352/lq.10112> (2 November, 2019).

Mons, B., ‘FAIR Data Publishing Group’, *FORCE11*, n.pag.
<https://www.force11.org/group/fairgroup> (2 November 2019).

__., ‘When privacy-bound research pays for open science,’ *EuroScientist Journal*, April 2016, n.pag. <https://www.euroscientist.com/privacy-bound-research-pays-open-science/> (2 November 2019).

Mons, B., et al, ‘Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud’, *Information Services & Use*, vol. 37, no. 1, Jan. 2017, *content.iospress.com*, pp. 49–56 <doi:10.3233/ISU-170824> (1 November 2019).

Sansone, S.A., et al., ‘FAIRsharing as a Community Approach to Standards, Repositories and Policies’, *Nature Biotechnology*, vol. 37, no. 4, Apr. 2019, pp. 358–67. *www.nature.com*, <doi:10.1038/s41587-019-0080-8> (30 October 2019).

Verhaar, P., Thysiana, (2020), GitHub repository, <https://github.com/peterverhaar/thysiana> (17 February 2020).

Weedon, A., ‘The Uses of Quantification,’ in S. Eliot and J. Rose (eds.), *A Companion to the History of the Book*, (Malden, MA: Wiley-Blackwell Publishing, 2009), pp. 33–49.

Wilkinson, M.D., et al., ‘The FAIR Guiding Principles for scientific data management and stewardship’, *Scientific Data,* 3, 2016, pp. 160018 <doi: 10.1038/sdata.2016.18 (2016)> (30 October 2019).

Websites Consulted:

http://contextus.net/ontology/ontomedia
http://experimental.worldcat.org/ontology/library/

http://id.loc.gov/ontologies/bibframe/

https://lov.linkeddata.es/dataset/lov/
http://purl.org/dc/terms/

http://purl.org/ontology/bibo/
http://purl.org/cerif/frapo/
http://schema.org/

https://w3id.org/spar/fabio