Epistemic Data Practices

Case Study report for the CWTS/Elsevier study on open data

[doi:10.17632/bwrnfb4bvh.1](http://dx.doi.org/10.17632/bwrnfb4bvh.1)

Clifford Tatum, Researcher

orcid.org/0000-0002-2212-3197

CWTS – Leiden University

Alex Rushforth, Researcher

orcid.org/0000-0003-3352-943X

CWTS – Leiden University

Thed van Leeuwen, Senior Researcher

orcid.org/0000-0001-7238-6289

CWTS – Leiden University

Sarah de Rijcke, Deputy Director

orcid.org/0000-0003-4652-0362CWTS – Leiden University

04 April 2017

Contents

Preface 2

Introduction 2

Case 1, Soil Mapping 3

Case 2, Human Genetics 7

Case 3, Digital Humanities 11

Discussion 15

References 16

Appendix 1 – Interview Schema 17

Appendix 2 – Information Sheet for Interviews 19

# Preface

This case study report is part of a larger research collaboration, focused on open data practices, between Leiden University’s Centre for Science and Technology Studies (CWTS) and Elsevier. A summary of this report was included in the main project report, Open Data: The researcher perspective, along with outcomes from two other complementary approaches: exploratory bibliographic analysis and an international survey of researchers.

# Introduction

Studies of open data often focus on the status and potential of making data publicly available for reuse by academic actors situated outside of the local data production context or by public actors not directly associated with academic research (Borgman 2012). This formulation of open data imagines the widest practical range of potential (re)users and invokes significant effort to prepare data for use by actors unknown to those who created the data. In this study, we shift the focus from concerns about making data publically accessible to the assessment of data sharing practices in the actual context of conducting research.

Three case studies are selected to provide contrasting views across different research contexts: Soil Mapping, Human Genetics, and Digital Humanities. Empirically, the study draws on interviews with key actors involved with data collection, analysis, and deposition, as well as on document analysis of materials such as (project) reports and websites. Conceptually, we adapt Leonelli’s (2013) framework for open data in the development of six epistemic dimensions associated with data practices: (a) situating data, (b) pragmatics of sharing/reuse, (c) incentives for sharing/reuse, (d) governance/accountability, (e) commodification and (f) globalization. Both the interview questions and subsequent coding scheme were developed on the basis of the dimensions (see Appendices 1 and 2). Following is a description of the dimensions.

1. Data situated: As the case studies were selected to provide a diversity of research contexts, we use the “data situated” dimension to explore contextual factors associated with data sharing in daily practice.
2. Pragmatics of data sharing and/or reuse: The activities involved with making data shareable require coordination of the tools, procedures, and standards, as well as communication among collaborators, which together enable transfer of datasets through different stages of the research process.
3. Incentives for sharing and/or reuse: Sharing data is often not rewarded in research assessment. This dimension focuses on both internal (e.g., collaboration) and external (e.g., policy) incentives that either promote or hinder data sharing.
4. Governance and accountability: With increased political attention to open data, we focus on the role of mandates, data management procedures, and training associated with preparing data for sharing.
5. Commodification: Data sharing and reuse often involve third party entities to provide services in support of data process, analysis, and storage. The aim of this dimension is to examine the role of licensing, commercial data services, and commercial funders in data sharing.
6. Globalization: Sharing and reuse of data within a local research team necessitates coordination among collaborators. International collaborations have the potential to further complicate the pragmatics of data sharing. The aim of this dimension is to examine data sharing practices that are distributed across international boundaries.

In the following sections, we outline each of the three case studies.

# Case 1, Soil Mapping

The site described in this case study is an international center dedicated to gathering information on world soil. Over a period of decades, the willingness to share data with the center by researchers from outside of the center has resulted in a host of data pertaining to soil properties of particular regions. The center receives donations, external grant funds, and block funds for projects mapping soil classes and properties of particular nations, regions, or across the entire globe. These maps are then used for modeling various ecosystem characteristics, such as climate change, soil erosion, soil nutrients, land use capacity, and soil biodiversity. Large-scale mapping initiatives tend to be the domain of dedicated centers, including national governmental agencies.

## Background

The field of soil mapping produces maps for modeling various ecosystem characteristics. These maps are used for monitoring features such as climate change, soil erosion, soil nutrients, land use capacity, and soil biodiversity. Although large-scale mapping (e.g. 1:5000 to 1:10.000) is frequently accomplished by individual researchers, large mapping initiatives tend to be the domain of dedicated centers, for instance National Governmental agencies. As such, given the scale of the phenomenon it deals with, the field has been in many ways ‘global’ already for many decades, with national and international governance initiatives reaching back to the 1950s and 1960s.

## Situating data

The first ‘inputs’ into the soil mapping process are soil surveys collected or shared by other scientists. These surveys provide both field information and information based on subsequent laboratory analysis (e.g. chemical and physical properties of soil sampled from the field). To produce the maps of soil types (classes) or soil properties, traditionally the soil mappers used mental models to draw spatial distributions of the soils, based on their interpretations of landscape maps. Nowadays many statistical and geostatistical methods are applied to predict the soil classes and properties for spatial prediction. To use soil survey data in a model, the soil descriptions have to be typed into a table (e.g. Excel), which can then be imported into software to run models. Through this approach, the scientists use soil information to predict maps (e.g. on soil erosion, soil fertility).

In the above description, the database appears as a central feature in the ‘data pipeline’ of the center. Further than this, when asked about meanings given to data, the database figured prominently in structuring informants’ definitions. As one informant explained:

*I would define it as systematized observations ... and what I mean by that is enough to know how that data came to be. Otherwise I don’t think you can really use it, you might say, “We have observations,” you don’t really have data... That’s why we speak of a database. It’s got structure. You know what every field is and what it stands for. That’s what I would call data.*

Strikingly, the very definition of data appears to draw on or anticipate sharing and re-use – capabilities which are achieved via systematization in databases. However, while definitions inter-linking data to databases appeared relatively straightforward and logical, it soon became clear that practices tying data and databases together could vary considerably across their field. In the following sections, we consider how informants relate to sometimes quite heterogeneous data sharing practices encountered in their day-to-day work.

## Pragmatics of sharing and reuse

In the center we studied, a small research team digitizes soil reports using a standard table format. As soil data was collected in different countries and at different periods of time (‘legacy data’), they tend to be based on different soil classification systems and different laboratory measurement methods. To be employed in a regional or global model harmonization is therefore needed. This process involves harmonizing data according to standards laid out in the United States Department of Agriculture and the National Cooperative Soil Survey and The World Reference Base (WRB) of the United Nations Food and Agriculture Organization (FAO)/UNESCO. Some surveys also lack metadata information (e.g. coordinates of soil location, information on the method), which requires ad hoc solutions within the teams.

When maps are produced, these are made freely available. The center also attempts to make the data on which the maps are based shareable via their web portal. For this the center has developed its own centralized and user–focused server database. However, we found that even in a mature data-sharing field like soil mapping, practices of data sharing are rarely straightforward. In the section Globalization, we shall see that re-sharing soil survey data is subject to different licensing controls imposed by the source from which the data has originated. Interestingly many of these restrictions emanate from national laws around privacy and data sharing.

## Governance and accountability

When questioned about general motivations for researchers to share data, informants offered several recognizable responses relating to the idea of data as a public good. Outcomes which could be gained through sharing and reusing data included altruism, ‘good karma’, reducing waste (through repeating work which has already been done), and providing a gift to the scientific community (and by extension societal impact from the work of the field). Other informants drew on an ideal of public good in stating that data sharing increases the chances that information is preserved rather than lost (through custodianship by the organization) and that the ‘well of knowledge and creativity’ will be increased via sharing:

*The more data you have, you can exchange, the more creative you can be. Other innovative processes can be stimulated that is the main reason also. With countries also like we have, for example had for Cambodia, you know, the history. And a lot of data has been lost so they were very happy that we know what we they have, safeguarded. We have got a copy of their data so they know at least it is…So for some organizations it is just to ensure that you safeguard their data.*

Part of the reason the ‘public good’ repertoire made sense in this case was that few obvious direct advantages could be gained by individuals in sharing data. For instance, sharing data was not seen as a source of reputational capital for scientists, in the same way as having one’s publications cited regularly by peers tends to be. Partly this was because using others’ ideas and others’ data were not seen as equivalent within the culture of the field. Despite data reuse not being a major currency of reputation, informants recognized certain standards were adhered to across the field with regards to citing data sources.

*Your name is mud if you don’t properly acknowledge data and there is a good standard because in the paper, you clearly say where the data came from. You have to say where the data came from, and acknowledge, “It was kindly provided by X or Y.” Or even if you just got it off the internet, you specifically say, “The dataset was obtained from the National Resource Conservation Service.” Everyone does this. I’ve never seen anyone in our field not acknowledge where their data came from.*

Although a minimum level of acknowledgement towards a source of data is expected, that source would more likely be mentioned in the main text or footnotes but seldom in the reference list or acknowledgments section. Likewise, the form in which a reference can be made is heterogeneous. It seemed then that the priority is for the reference simply to be made, not for a particular referencing style to occur. Practices of acknowledging data re-use are not necessarily taught through formal training or ‘awareness’ courses, or communicated via policy instruments, but simply get introduced to young researchers entering the field in a similar way as students would learn to cite other publications through reading and submitting written work. When sharing data on which the maps were based, the center provided instructions on how their data should be cited. While users tick a box agreeing to this as part of the terms and conditions on the web portal, the informants recognized they had no way of policing this and that indeed few adhered to the correct citation format in practice. From the perspective of their own accountability as a center, this lack of consistent data citing practices meant they could not be measured by committees accrediting them according to number of times their data had been reused in publications.

However, while they are not made accountable through data citations, the web portal enables them to collect information on those who download the data, which can be used both for strategic monitoring and external accountability:

*… really the key thing is that you want to know what the impact of your data is. That’s very interesting to us of course. Of course, we want to make data that is more used, especially with our target users or with our clients. We want to be able to trace that. We do of course trace the… You know today, if you have a web service, you can trace where the people are coming from, how much time they spent, how much megabytes they download. You can even trace when they are on the website, you can trace their mouse, you can see what they are looking at.*

Although the ‘impact’ of research data is not (yet) measured of individual scientists, the role of the center and its accountability towards public funding sources means ability to show the impact of their data sharing platforms has become part-and-parcel of being an accountable body. This case thus provides an interesting example of how data sharing has become an activity for which certain kinds of organizations can be audited as service providers.

## Globalization

As mentioned, the field of soil mapping has long been ‘global’ in its intended reach, and sharing across national and institutional borders goes back many decades. In principle, the research mission of the field and goals of sponsors tend to align quite well around this ideal. Yet this does not explain why many kinds of actors in countries worldwide are willing to ‘donate’ or make available their soil survey data to the center. When probed on why governmental organizations were willing to share data with them, the informants cited various important explanations: official policies of governmental organizations to share data; the lack of a commercial threat in allowing others to use their data; the possible benefits their own countries or regions might gain from having their soil landscapes mapped; not being concerned with gaining priority for discoveries (which can restrict the sharing of data among researchers). However, there were also some frictions encountered in trying to obtain data from countries or for the center to try to share data they have obtained from elsewhere. Friction between national practices was a common repertoire in explaining blockages. One example was the diverse national and regional laws surrounding data privacy and licensing in some countries.

*The point data may have privacy issues and France is a good example there because of the law in France. This is general for everyone, you can’t give identifiable data. It’s got to do with identifiable data also for any kind of survey. My French colleagues can use the data to make their own maps but they can’t give out the points…*

*In some countries, there is actually a restriction on the scale you can show. China, for example, they will give you the polygon maps that are made up to 1 to 1,000,000 scale… I can’t remember is it’s 1 to 1,000,000 scale or 1 to 250,000, I can’t remember but the detailed country maps, which they’ve made 1 to 50,000 or better, this is by Chinese law, cannot be given out.*

Also cited were diverging ‘cultural’ assumptions about the terms under which data should be exchanged (e.g. should monetary exchange occur, or should data sharing be understood as part of a ‘gift economy’).

*Other times it’s purely for financial reasons and that’s the case with the British soil survey. I should say England and Wales, not all of Britain. The Dutch soil survey is similar, they are much nicer about it. They have a fee, you have to buy it. It actually is open but you have to buy it.*

Whereas some countries had fixed prices attached to accessing data, in the experience of one data officer, negotiations with actors in some Western African countries sometimes led to misunderstandings, for instance an expectation of informer ‘barter’ type exchange processes:

*Like many countries when I talk to them, “Yes we want to work with you; we have a lot of data.” But the moment you really are approaching the point, there is no data or maybe there is data but only if you make a deal. So they are not ready to share it, they only want to, let’s say, sell.*

Whereas this example hinges on a lack of bureaucratic protocol in obtaining or exchanging data, at an opposite extreme the process of sharing could be caught up in a bureaucratic infrastructure:

*The whole Indian government is very bureaucratic from top to bottom and in a bureaucratic organization, a lower level employee is scared silly of making a decision of any kind. They get higher and higher and higher and higher and to get data from India, you practically have to get it from the Prime Minister. Everyone below, even if it would be allowed, is scared to make the decision.*

In sum, the social, financial, cultural, and legal contexts of data sharing vary considerably across countries, making exchange of data for the soil mapping center far from seamless or routine. Working as a node in a ‘mature’ data sharing field, the accounts of informants from this center suggested that while data can in principle be ‘open’, obtaining data can be a laborious, uncertain process. Likewise, differing bureaucratic structures across countries and regions posed particular challenges with respect to establishing common standards for data re-use and sharing.

## Commodification

For the soil mapping case, the commodification dimension is closely related to diverse national and regional differences. As noted above, some friction is encountered in the exchange of data across different international contexts. With the proprietary roles and involvement of non-academic actors come financial costs attached to data in certain national contexts. This is likely to frustrate any realistic chance of the field becoming entirely free and open in its data practices any time soon.

# Case 2, Human Genetics

The research center selected for this case study is organized into several co-located biomedical genetics labs. A centralized bioinformatics group provides data processing and analysis expertise to multiple labs in the research center, coordinating their activities with several projects simultaneously. We focus on bioinformatics data sharing practices in clinical genetics research on rare diseases.

## Background

Data associated with human genetics involves a complex series of processes, typically beginning with a blood sample, from which genetic material is extracted and then sequenced in a computerized process, referred to as a sequencing machine. The output format is a string of text using the letters representing the four possible nucleotide chemicals: G, A, C, and T for DNA, and G, A, C, and U for RNA. Letter clusters are arranged as a long string of text without gaps and stored in a digital format.

In this case study, modes of data sharing are organized by local collaborations necessary for diagnosing rare diseases and in some instances in collaboration with other clinics also investigating rare diseases. The local collaborations are typically comprised of a clinical geneticist, a wet-lab technician, a bioinformatician needed to prepare and analyze the data, and often additional researchers in the same center or in other clinics to assist with interpreting the outcomes from the data analysis.

## Situating data

In the genetics lab, collaborations typically consist of multiple people, each with particular expertise needed to fulfil tasks. The mode of sharing data is defined by the need to locally share or to transfer the data between the different stages of data processing, analysis, and interpretation. Informants characterized data in this context as digital or digitized versions of the genetic source material. As one informant stated:

*Data is, for me, any form of information. You can make it very philosophical, or you can take the physics approach. So, data can be really a lot to me, but in my daily work, it is everything that can be digitalized, in a sense. So, even pictures, for me, are data, as long as you can just scan it in and put it in a report [...] But of course, the data that we work with is usually stored in databases [...] a very long list of variants, that of course also need to be shared.*

They further describe data in relation to a database; datasets are stored and processed in databases throughout the course of the research project.

*But of course, the data that we work with is usually stored in databases, so I don’t deal with text that much, but more with a very long list of variants, that of course also need to be shared.*

Transport of data over the course of a research project creates an apparent epistemic distance between source material and the object of analysis. Throughout the process, metadata is added to a dataset at each stage of data processing. As a consequence, metadata take on increased importance in the course of the research, both as the object of analysis and for enabling distribution of datasets among collaborators.

## Pragmatics of sharing and reuse

The bioinformatics group provides data processing and analysis expertise to multiple labs in this research institute. In this central role, the group coordinates data processing and analysis among several projects simultaneously. As such, sharing and reuse of data is integral to the research objectives and is embedded in the research design. The data tasks typically begin with a digital version of sequence data, which then undergoes many layers of analysis according to the intended research inquiry.

*After sequencing, we compare it to what we think is a healthy reference, [...] I find variants in genes that are responsible for muscle function, [...] report that in principle back to the patient, then that part is done.*

In most cases, this involves analysis in a semi-automated bundle of routines referred to as the ‘pipeline’.

*[The] raw reads that come out of the sequencer [...] are transferred to a specific location on [the server] cluster, and then automatically this pipeline starts processing this data. The pipeline consists of tools that first take these reads through some quality control, so I still have reads, and in the next step I go, from those reads, I align them, so then I have aligned reads, but not variants yet. So, then I take that, I enter it into a variance caller, then I have a list of variants; but they are not annotated yet. So, it’s just, yes, lots and lots of very small, sequential steps to come to an end product.*

Bioinformaticians are tasked with developing and maintaining the suite of analytical routines while at the same time pursuing increased automation of the pipeline. While advances in sequencing genetic information did bring about a standardized and somewhat routine process, intricacies of the pipeline analyses still involves on-going development of customized analysis procedures. In this way, the pragmatics of data sharing and reuse are included in the translation of the biological question into an appropriate data model. In the case of rare disease research, the data model facilitates both local reuse and external sharing, while at the same time following strict privacy protocols associated with human genetics.

*[W]e store all of the observations in our system, so that whenever we find something again, we can either come to the same conclusion or not, [...] A human doesn’t need to look at that anymore. [We then] store all of the variants in a big database [...] it will only answer in frequencies, [...] you cannot do any queries on the individual level, because asking, [...] I could identify a person; but by just asking frequency information, I still don’t know anything except whether or not a variant is rare or frequent in a population. When it is rare, it is a higher chance of being pathogenic.*

Informants articulate the skill set needed for this function as a combination of the capacity to provide consultancy; work with the genetics researchers to understand the specific requirements of data for the ‘biological question’ and for data modelling; and design and implement a data model as a component of the research method.

## Governance and accountability

Informants report that funding mandates do not sufficient motivate follow-through regarding publication of data in a repository. Instead, publishing mandates were identified as being more effective.

*Funding agencies, they now start to impose [open data mandates], but they do not control whether you’ve really... they do not check whether you’ve done it, right? So, there is not a penalty for this. [When publishing identification of a rare disease] it is mandatory to publish the [genetic] variant., if you want to publish with a couple of these journals, they say you have to put those variants into a database.*

There are also significant privacy concerns when sharing genetic data. Once a particular genetic variant is identified and the rare disease has been diagnosed, the data are stored for local reuse and made available for sharing through a linked network of other rare decease databases. To address the tension between the medical benefits of sharing genetic data and privacy concerns, only the frequency of genetic variants is shared. Privacy and security concerns are managed through sharing only the information needed to help others diagnose rare diseases without exposing genetic data that could be used to identify individuals. Informants indicate that while it is not common, it is nevertheless possible to identify individuals from genetic data.

*In principle, studies have shown that you can identify an individual based on just his genome sequence. So, that’s why there are privacy concerns […] a patient has a kind of condition which is quite rare, and [he or she] lives probably in the area around [the rare disease clinic], because that’s where the sequencing was done. And when you collect all the data together, then in principle, yes it could be possible [to identify an individual].*

Given these privacy and security concerns, the sharing of rare disease data is configured to provide ‘frequency’ queries without providing access to the genetic data itself. For example, an external researcher can query the database for a genetic variant, which will return the number of known instances of the variant. If the number is high, the variant in question is not likely to be associated with a rare disease. If, however, the variant is rare in the database, it is likely to be pathogenic. Moreover, the database is central not only to sharing and reuse of genetic variants, but also in administering rigorous security measures to comply with data security protocols.

## Incentives for sharing

Like the other dimensions considered above, incentives for sharing are largely embedded in how the genetics research is organized. With the organization of data processing and analysis as a centralized function, transfer of data among collaborators is a necessity for conducting research. Informants indicate benefits for the medical fields in general and for patients in particular as incentives for sharing data beyond the boundaries of a particular project. However, they also experience resistance to sharing.

*Politics. It’s purely politics. Everyone always thinks it’s a good idea, but when you say “Okay, now, come send your data, we’ll put it in this database.” Then people always have concerns. “Yes, I’m not sure I can do this. I need to have signatures from this, or those persons.” And they never get it, and you always end up with long, long discussions why they can’t share it. They analyzed it in a slightly different way, or I don’t know, you always get tons of excuses. You never get an answer, or they just say “It’s just too much effort for me.”*

The example of sharing frequency data for rare diseases, outlined above, illustrates a particular mode of sharing that has been successful. However, in spite of datasets already conditioned for sharing within a collaborative project, there are still additional tasks needed for sharing data externally. Other barriers to sharing identified by informants include lack of professional credit for making the datasets sharable, additional time and expertise needed, and resistance from others to share their data.

## Globalization

The dimension of globalization did not seem particularly relevant for this case. Although present in the human genetics projects, data practices associated with globalization appeared to be embedded in the normal course of research collaborations. For example, international research collaborations are common, as is the distribution of tasks among participants on the basis of differentiated expertise across international collaborators. These collaboration practices do not appear to create additional frictions with respect to data sharing.

## Commodification

In this case study, commodification is present in the form of third party commercial services, but the rare disease research did not seem to be effected by it. Informants did however identify commercial tensions with third party contributions to genetics research in general.

*I think there is a lot of tension there, because often, informed consents from patients are not clear on this, whether their data can be shared with industry. Industries themselves do not want to share at all. They are very protective of their own data, so it’s not often a real mutual thing […] NDAs (non-disclosure agreements) are always very painful to come up with. [Commercial partners] are worried about their intellectual property, obviously, and indeed from the hospital, or patients, they’re a little bit worried that a company will make profits from their data.*

As third party commercial services, such as gene sequencing, are integral to genetics research, data collection, and data practices associated with commodification appear to also be embedded in the normal course of research collaborations.

# Case 3, Digital Humanities

## Introduction

Many digital humanities research projects in the Netherlands are linked through a national level network. For this case study, we focus on researchers whose work straddles the traditional humanities and computational science. An important part of sharing in this context is the ability to transfer data among project participants, which enables later reuse (for instance, to extend or compare with existing data). What is shared is the data itself, along with the analysis and processing tools. In addition, transparency and reproducibility are important, but these can be very expensive to implement and researchers are not incentivized to do so. While this is a significant issue for the field of computational science, we will see that it is less so in the humanities.

## Background

This case study is focused on various projects in which analysis of texts figures prominently. In this context, data typically begins as a corpus of digitized texts. An important part of sharing in these projects is transferring data among project participants, which enables sequential analysis across a defined research process.

## Data situated

The composition of project participants and the distribution of research labor varies across digital humanities projects. Among projects there is often an ongoing tension between converging fields and their respective communities of practice. Sometimes more senior humanities scholars do not have an interest in what they refer to as 'technological' issues and prefer 'intellectual' work.

*Humanities is a shared environment, we share a lot and we think that we all improve from adding new layers on top of each other instead of just doing all the same things differently, all different spots.*

It appears there are not very many cross-overs in this regard. It is mainly PhD students who realize they need to develop these skills or who set out to develop expertise across domains.

*But most of us started at the hardest point, the last point, thinking about how can we digitize our very complex questions about what is democracy, how did it develop over time, and can we find that back in text. In order to do that, you need to learn a completely different kind of trade. For instance, I’ve learnt to make an annotation system myself for finding these emotions.*

Regarding the concept of data in this cross-disciplinary context, informants identify a process of digitization that turns objects into data.

*What is data? As a historian, I would say those are objects that are turned into information by a digitization process. For me they're historic objects, objects that can be part of a museum collection and they're still objects. At the moment, they are digitized and people choose what information layers to add to those objects in a digital environment, then they turn into data more than into the historical archival works that they have been.*

Data gathering and processing involves computer scientists who collect 'raw' data from the web, write scripts, and store the data locally on a server that 10-12 people can access. At some point the pre-processed data set is 'frozen' if it is used for research. It is usually in a database and data is usually shared with the results of each preceding analysis along with and any software tools that were used or developed.

## Pragmatics of data sharing

Sharing and reuse of data in digital humanities is generally bounded by the particular collaboration. Humanists, information scientists, and computer scientists, for example, work together to analyze traditional humanities research objects with the benefits of digitized content and computational methods.

*If you’re not willing to collaborate, you might as well stay far away from digital humanities because that won’t work. There are always exceptions, the genius who can combine it all, but the risk of overlooking vital things is very big, I think. So, I wanted to have a team with overlapping expertise.*

An important part of sharing is transferring data among project participants, which enables later reuse (for instance in doing an extension on, or comparison with, existing data). What is shared is the data itself, along with the analysis and processing tools that were used in and developed for a particular project. Meta-data is crucial for sharing among project participants; for understanding, interpreting, reusing the data.

*I think metadata is always data in the way that it is consciously added as a layer of digital information, specifically for a digital context […] If you scan that text, OCR it, and open it up through the web, then it turns into data, although you didn't change a letter and it’s still completely the same 19th century edition of this mediaeval manuscript and you start calling it data at the moment, it is presented.*

Datasets are usually stored in a database or as a big file that can be zipped and they are typically stored locally. Others can request access to datasets, which are then provided via a web server.

## Incentives for Sharing

Obstacles to data sharing in the digital humanities case include tensions in the distribution of labor and publications not specifying which processing tools, or which version of the datasets were used. One of the main problems is that while most informants value sharing, the career benefits in doing so are uncertain.

*I think that thing about the funding, and the recognition that this is also work […] and it's work that needs to be valued. You know, for most researchers, this is really boring, unglamorous work, for which you get no credit. you do not get credit for producing a really nice dataset that thousands of researchers around the world can use to do interesting work.*

There are indications that this is also changing, with funding agencies introducing preferences (if not mandates) regarding data management and transparency. However, the primary incentive for sharing is the need to transfer data across different stages of data analysis within a cross-disciplinary context.

## Governance and Accountability

Training related to open data was generally understood as beneficial and/or desired, but it appeared to be largely missing. Instead, collaboration with computer science and information science serves as an important dimension for establishing the skill-sets required for data-intensive research.

*[A computer scientist on the team] tried to teach the others a little bit of how he worked, working with a Python Notebook and things like that. That proved to be too difficult, too complex for the others. Something else we did learn, was to use R and then the [writing] style package, the package for which we do stylometric analysis […] if you’re a literary scholar, you’ve not learned to do any statistics at all, or any programming at all. I hope that will change in the future, but we have to deal with what we now have.*

In addition, Humanities scholars identify use of GitHub, which is commonly used by computer scientist to share software, as an important means for sharing data. This illustrates the transfer of highly specific practices between disciplines and the utilization of resources across particular types of collaborators, rather than following typical repository-oriented resources associated with the open data movement more broadly.

## Globalization

Intellectual property rights (IPR) is an issue of particular concern in this field. Sharing data, for example, often involves addressing licensing of literary texts. Data is relatively easy to access with the use of web scraping tools and techniques, but it is hard to penetrate commercial and legal attributes related to platform owners like Amazon or Library Thing, as well as content owners such as university libraries or content publishers. However, there appear to be significant cultural differences in the ways in which cultural objects are treated.

*In the Netherlands, our libraries and archives are pretty flexible in providing information. I was just in Japan which has a completely different idea about for instance museums as treasure holds. They protect the treasures of culture and they would never consider opening that up just freely for the public*.

The international links that came to the fore in this case study were primarily in Europe and North America. The Japan example suggests that the further research would be needed to better understand the globalization dimension in Digital Humanities.

## Commodification

As a new/hybrid field, there is also variability in how commodification is addressed by digital humanities. There is increasing awareness of commercial opportunities, and presence of commercial parties at conferences. We found instances of both enthusiasm and skepticism associated with commercial partners. Some welcome the opportunity to partner with commercial third parties with an eye toward long-term development and sustainability of the humanities discipline. An interesting example involves the development of text analysis software with a commercial entity.

*A small company, a consultancy company who works on projects for publishers wanted to know if they could use our corpus, because they were trying to predict a best-seller. We got together and talked and found that we had a lot of interesting things to share and to discuss. So, now we’re working on a new project in which we try to develop a scouting tool for publishers. [to fund this project] we did a pre-proposal and we’re waiting to hear if we can develop a tool for proposal. [...] That’s a call that I’ve never submitted something to before, so that’s also very new and I like this kind of collaboration. It’s valorization to use a Dutch word, societal relevance is what interests me.*

For others, commodification is still a dirty word. But this is also changing as a result of priority setting by funding agencies (e.g. the requirement of bringing in the creative industry). So, commodification is happening, but only at a very slow pace. For example, start-ups cannot provide matching funds that are often required to participate, which is also a barrier.

# Discussion

All three cases reveal ways in which the pragmatics of data sharing and reuse are embedded both in conceptions of data and in normal data processing work. Data is quite often described as digital, structured, and in relation to databases. Observations or source materials become data upon deposit in a database, which renders data accessible for sharing and further processing. A particular dataset thus constitutes digitized source material (e.g. sequenced genes, old texts, or soil samples) and layers of metadata that result from each stage of the analysis. A digitized text, for example, could include embedded information from optical character recognition, sentence structure analysis, grammatical structure, syntax analysis, and style analysis. Each procedure may be conducted individually or by different people (e.g. digital humanities), or is bundled in an automated process (e.g. genetics pipeline).

In the final stages of data processing, metadata is of particular interest and source materials less so. The database is thus integral to data analysis routines and to sharing among collaborators who participate in different stages of analysis. There is an apparent epistemic distance between source material and what is considered data. Together with the diversity of data practices across cases, this raises questions about the efficacy of policy that prescribes open data as an activity detached from situated contexts. [*relevant dimensions: data situated and pragmatics of data sharing*]

Material from the Soil Map case illustrates potential globalization challenges regarding ‘late stage’ data sharing and reuse practices. Friction from national differences was evident, including diverse national and regional laws surrounding data privacy and licensing. Also striking were diverging ‘cultural’ assumptions about the terms under which data could be exchanged (e.g. should monetary exchange occur, or should data sharing be understood as part of a ‘gift economy’). In addition, differing bureaucratic structures across countries and regions posed particular challenges with respect to freeing-up data for re-use and sharing. [relevant dimensions: globalization and commodification]

In all the case studies, sharing and reuse of data is shaped significantly by collaborations. Incentives for sharing and reuse are embedded in a research design that involves distribution of expertise across collaborators. This involves organizing a collaboration on the basis of multiple participants needed to fulfil necessary tasks in a particular research project. As such, this mode of sharing is shaped by the need to share or transfer the data among collaborators across the different stages of data processing, analysis, and interpretation. A notable exception is an account of compliance with Open Data mandates from journal publishers. In this instance, researchers were required to deposit data as a condition of publishing the article. [*relevant dimensions: incentives for sharing and governance/accountability*]

Training related to open data was generally understood as beneficial and/or desired, but largely missing. In the Digital Humanities case, for instance, collaboration with the computer sciences and information sciences is an important dimension for establishing the skill-sets for data-intensive research. There is some indication of transfer of skills and practices between disciplines in the development of data sharing practices (e.g. use of GitHub by Humanities scholars). [relevant dimensions: governance/accountability and pragmatics of data sharing]

# References

Borgman, Christine L. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63, no. 6 (2012): 1059–78. doi:10.1002/asi.22634.

Leonelli, Sabina. “Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology.” *Bulletin of Science, Technology & Society* 33, no. 1–2 (February 1, 2013): 6–11. doi:10.1177/0270467613496768.

# Appendix 1 – Interview Schema

|  |
| --- |
| ../../../../../../../../../../var/folders/ng/rmm767k967d5njfxzjqm94nw0000gn/T/com.apple.Preview/com.apple.Preview.PasteboardItems/interview |
| ../../../../../../../../../../var/folders/ng/rmm767k967d5njfxzjqm94nw0000gn/T/com.apple.Preview/com.apple.Preview.PasteboardItems/interview%2 |

# Appendix 2 – Information Sheet for Interviews

|  |
| --- |
| ../../../../../../../../../../var/folders/ng/rmm767k967d5njfxzjqm94nw0000gn/T/com.apple.Preview/com.apple.Preview.PasteboardItems/CWTS%20inf |
| ../../../../../../../../../../var/folders/ng/rmm767k967d5njfxzjqm94nw0000gn/T/com.apple.Preview/com.apple.Preview.PasteboardItems/CWTS%20infos |