universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

The Valuation of Data through Biodata Resources Examining Registers of Valuing and Valuation Practices in and around Biological Research Data Infrastructures

verfasst von | submitted by

Roman Michael Hansen

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 906
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Science-Technology-Society
Betreut von | Supervisor:	Univ.-Prof. Sarah Davies BSc MSc PhD

Abstract – English

Biological data and knowledge bases are essential infrastructures in the life sciences, supporting both experimental research and computational analysis. More than repositories, these resources actively structure scientific knowledge through processes of selection, annotation, and integration into wider data models and standards. Such work is not purely technical but deeply social, involving, among many, curators, maintainers, and users, whose practices shape and are shaped by the design and governance of the resources.

This thesis investigates how data are evaluated and valorized differently as they move through these socio-technical systems, and how such valuations vary across positions and practices. Drawing on interviews with maintainers, curators, and users of small to medium-sized, often community-driven biological databases, the analysis identifies key registers of valuing, including epistemic qualities and questions of usefulness and usability, as well as non-epistemic, personal and communal benefits and ideals of openness and control. Building on these registers of valuing, practices involved in working within, interacting with, and maintaining resources are traced out, highlighting conflicts that arise and negotiations that need to be made.

Throughout the analysis, maintainers emerge as central figures whose interdisciplinary expertise enables them to bridge biology, information infrastructure, and community engagement, while also navigating field-wide collaboration and the expectations of institutional and funding stakeholders.

By bringing valuation studies into dialogue with scholarship on data-work and maintenance, this thesis offers a rich account of the hidden, skilled, and affective labor underpinning biological data infrastructures. Putting into question descriptions of maintenance as mundane work, this thesis highlights the complex negotiations through which data, and the resources that hold them, are valued.

Abstract – German

Biologische Daten- und Wissensbanken sind zentrale Infrastrukturen der Lebenswissenschaften und unterstützen sowohl experimentelle Forschung als auch computergestützte Analysen. Mehr als bloße Repositorien strukturieren diese Ressourcen wissenschaftliches Wissen aktiv durch Prozesse der Auswahl, Annotation und Integration in umfassendere Datenmodelle und Standards. Diese Arbeit ist nicht nur technischer, sondern in hohem Maße auch sozialer Natur: Sie umfasst Kurator:innen, Maintainer (frei übersetzt Instandhalter:innen) und Nutzer:innen, deren Praktiken die Gestaltung und Verwaltung der Ressourcen prägen – und zugleich von ihnen geprägt werden.

Die vorliegende Arbeit untersucht, wie Daten innerhalb dieser sozio-technischen Systeme unterschiedlich bewertet und valorisiert werden und wie sich diese Wertungen über Rollen und Praktiken hinweg unterscheiden. Auf Grundlage von Interviews mit Maintainer:innen, Kurator:innen und Nutzer:innen kleiner bis mittelgroßer, häufig community-getragener biologischer Datenbanken identifiziert die Analyse zentrale Register des Wertens – darunter epistemische Qualitäten, Fragen der Nützlichkeit und Nutzbarkeit sowie nicht-epistemische, persönliche und gemeinschaftliche Vorteile sowie Ideale von Offenheit und Kontrolle. Aufbauend auf diesen Registern werden Praktiken des Arbeitens mit, des Interagierens mit und des Instandhaltens von Ressourcen nachgezeichnet, wobei entstehende Konflikte und notwendige Aushandlungen herausgearbeitet werden.

Im Verlauf der Analyse treten Maintainer:innen als zentrale Figuren hervor, deren interdisziplinäre Expertise es ihnen ermöglicht, Biologie, Informationsinfrastrukturen und Community-Engagement zu verbinden – während sie zugleich feldweite Kooperationen sowie die Erwartungen institutioneller und finanzieller Stakeholder navigieren.

Durch die Verbindung von Valuation-Studies-Ansätzen mit der Forschung zu Datenarbeit und Maintenance bietet diese Arbeit eine differenzierte Darstellung der verborgenen, hochqualifizierten und affektiven Arbeit, die biologischen Dateninfrastrukturen zugrunde liegt. Sie hinterfragt die Charakterisierung von Maintenance als alltägliche Tätigkeit und hebt stattdessen die komplexen Aushandlungen hervor, durch die Daten – und die Ressourcen, die sie zusammentragen – wertvoll werden.

Table of Contents

Acknowledgements

This thesis was developed as part of my employment with the *Technosciences, Materiality, & Digital Cultures* research group. It was here that my engagement with the topic of biocuration began and my interest in data infrastructures took shape. Being part of this group not only provided the financial support necessary to complete this project within a year but also offered an inspiring environment: a space to test ideas, a community to learn from, and a source of academic confidence.

I would like to thank Sarah Davies for entrusting me with the role of research assistant, for her guidance throughout this project, and for supporting my path into academia. Her ability to create an open, encouraging, and always-friendly working environment made the work on this thesis and beyond consistently rewarding. I am also grateful to my colleagues in the research group, who enriched my time in the department through their collaboration, generosity, and enthusiasm.

More broadly, I thank the STS department, its leadership, lecturers, and support staff for providing me with an academic home over the last two years. Here, I learned to approach science and technology with new critical perspectives. Most importantly, through coursework and the work on this thesis, I learned to write more clearly, more freely, and, hopefully, more enjoyably.

I also wish to thank my interviewees, who welcomed me into their world with generosity, patience, and expertise. They shared their time and stories with openness and honesty, making this project possible. I am equally grateful to Constantin Holmer, whose earlier work with Sarah Davies on biocuration offered me a valuable entry point into the topic.

On a more personal note, I am deeply grateful to those who made my first two years in Vienna pass so quickly and joyfully. Friends, old and new, helped me build a sense of home in this city. During the writing of this thesis, they provided both the necessary distractions and accepted my seemingly endless working phases. I am especially grateful to Alix for countless shared library sessions, and to Paloeka for providing the last fresh pair of eyes.

Finally, my deepest thanks go to my partner for constant support, patience, and calm throughout this process. To my family, and to those who like it: thank you for always being by my side.

1. Introduction

The importance of data in science is not new. Yet, its promises and regimes are amplified in an increasingly digitized research environment. Similarly, biology is continuously developing its digital side. The organic and living are more and more often captured, sequenced, and digitized, standardized, accumulated, and exchanged – moving across the world quicker than any of them ever could in their physical forms. In digital collections, genes and proteins, their interactions, and mutations are cataloged and distributed, findable for researchers from all life science subfields across the globe and throughout academic hierarchies.

Systematized in biodata resources that range from large-scale data repositories and narrowly focused knowledge bases, researchers can look for and find academic papers, sequenced genes, and experimental data. These resources, the data and knowledge provided, allow researcher to answer their questions *in silico* (meaning digitally and not in a lab) or to structure their further research. Now, an even wider range of researchers has access to data that was previously reserved only for a few, shortening the time required to assemble relevant knowledge and extending research collaboration across previously hindering boundaries – so at least the promise. Not all data are equally useful, and finding the right set of data is not without prerequisites; as data move between their context of creation and potential contexts of use, they need to be made movable (Leonelli, 2016, p. 70).

The aim of this thesis is to examine these processes of making data movable – to understand what they entail, who is involved and what is deemed important. Focusing on collections that provide specialized aggregations of research data and biological knowledge, I will explore questions of value and types of practices involved, both directly and indirectly. This means considering close interactions with the data, as well as the structural work of developing and maintaining digital infrastructures. I follow values throughout complex constellations to provide a better understanding of how the work necessary, and how the decisions required to make data movable, shape, and are shaped by their context. Ultimately, this thesis answers (RQ) *how data are valued in and around biodata resources*, (SQ1) *how they are valued differently through practice*, and (SQ2) *how different registers of value are combined and negotiated by those involved*. The resulting analysis will highlight different ways that data and data work are seen and enacted as valuable. Different types of data work will be systematized, and especially the importance and features of maintenance tasks will be laid out.

The collections centered in this thesis collect research on certain topics and provide their users with as much of the available information as possible in their respective areas of interest. That means the research that is published is embedded and integrated in a larger overview of the research context and adjacent fields. To achieve this integration, the studied organisms, the experiments conducted, and the measurements collected need to be described in a way that makes them understandable and relevant to more than just their creators. New research is standardized in a way that even in collections far away from its origin, data still remain tied to the realities they capture and are usable for new academic endeavors. However, a lot needs to happen before researchers are able to type, for example, their "favorite" protein into a search bar to get all the existing scientific knowledge and data on it. All

research on the protein needs to be described using the same term for this protein – ideally, a name that everyone agrees upon. Only after research is identified, described in standardized ways and published as part of a collection can it be found easily. If data and their descriptions are not compatible, if they remain unfindable, then what value can they actually provide to a global scientific community?

To make the movement of data more compatible, digital collections provide standardized forms of movement. Described as "databases" or "biodata resources" more generally, there are different types of collections to consider. Research repositories, collect experimental data, and research publications in large amounts. Especially in big repositories, the research collected is described in a standardized way, which serves a large community but might not provide all the details needed for certain sub-fields. Serving those fields, smaller resources strike a different balance in their curation efforts. Selecting a narrower sample while providing more detailed descriptions, they are more closely aligned to their users' needs.

Especially knowledge bases, which make up a substantial part of the resources considered in this thesis, follow this approach. Here, research is collected and described to provide more than an overview of the research field, but to provide an increasingly complete representation of a biological entity of interest. The results are harmonized and specialized collections that represent a subset of biology and do so in an interactive way. The resources considered in this thesis mostly do so on a small scale, focusing, for example, on a single organism and all the knowledge there is about it, or on certain use-cases, supporting cancer research or lab experimentation. Such knowledge bases provide their users with information about relevant entities bundled into compact pages. On such a page, a certain protein might be described in a number of relevant, yet also standardized ways. Each description is appended with a link to the original research providing the finding, ensuring that users can always access more information about the origin of the knowledge they receive (Leonelli, 2009, pp. 5-6). But these metadata do not emerge on their own.

As part of curation processes, automation is increasingly explored; however, the selection, description and annotation, as well as the publishing, are still a very human process. These curatorial efforts, also called "biocuration", bridge the gap between research in its original form and the structured aggregate. This is an important gap to bridge. As research papers and data published are not necessarily produced using commonly shared terms, everyone wanting to use them in large quantities is required to translate what they contain. Automatic systems might assist this process, yet to this day, a lot of this data work is performed by humans (Chen et al., 2022, p. 3). The biocurators performing it are highly educated, often with PhDs and bench lab experience in the life sciences (Davies & Holmer, 2024). Their task is to make biological research data findable, accessible, interoperable, and reusable (FAIR; Wilkinson et al., 2016). That means mapping data and knowledge to standardized vocabularies, integrating them with connected papers and datasets, all to allow future researchers to find all the data that might help answer their questions. The integration into shared resources is not just a means of distributing, but as much a means of sharing the work needed to make data movable

(Drysdale et al., 2020, p. 2636). By becoming movable, these data can overcome great distances between research fields, epistemic cultures, as well as time and space (Borgman & Groth, 2025).

The standardized vocabularies used to ensure such movability, also called ontologies, are ordering systems, defining, relating and delimiting entities of interest (Rubin et al., 2006). They are sets of possible connections between the subjects of any research, mapping out hierarchies and dependencies of biological entities. In biology, such systems allow for the description of elements like genes or proteins, and processes or interactions alike, promising to overcome the challenges described above (International Society for Biocuration, 2018). Pathogens, mutations, and gene markers can all be described using ontologies.

However, ontologies are not all alike, and ontologies are not always stable (Leonelli et al., 2011, p. 6). They shift and change according to the research they are embedded in and the community they serve. This is where the project of collecting all biological data gains its complexity – as if its sheer size was not enough. Across ontologies, the same process or entity can be described very differently, and these descriptions are not always compatible. The data journeys from the creation to the use as scientific evidence are not always straightforward, making the data's value uncertain at times.

When a protein is researched, possibly sequenced, and published, the information about it might end up in a database. Before it is displayed and findable, the automated system of the database tries to identify the researched protein and add metadata to help identify it. As part of the curation, the new entry is then mapped, often manually, as accurately as possible to relevant ontologies, like the Gene Ontology (GO), placing it in a larger research context (Chen et al., 2020, p. 98). Additional resources, e.g., describing the protein's interactions, are connected automatically based on the GO term, making the new entry even richer. These metadata are crucial to make data movable, as they describe how a derived piece of knowledge can be or should be interpreted. But these vocabularies only describe what is known. As new proteins are identified, existing ontologies might need to be extended (Leonelli et al., 2011). That means adding to the catalog of definitions, drawing new delineations, and mapping new connections. With the evolution of research, ontologies need to evolve as well. They are changing, constantly.

Biodata resources, their user interfaces and curation platforms provide the social and technical context for the data work described. Here, a space is provided for data of some kind, and here ontologies are built into the design (Nadim, 2016). Filled by curators and users but fundamentally shaped by those building and maintaining the resource itself, this space provides a constantly evolving context for data to become permanently useful. Constantly changing, yet built for longevity. These contrasts will become a focal point of this thesis. Looking at value and practices around biodata resources means looking at work that is done both directly with the data and on a more structural level. The result is a highly embedded way to see data movement.

Specializing in specific types or subjects of data, the resources discussed in this thesis provide visualization tools, useful data bundling and easy interaction that serve their respective research

communities and their data needs. Selection, curation and communication of data and knowledge are optimized for the research that they will most likely be used for. They are shaped by contextual needs but, similarly, are vital to how contemporary bio research is carried out. Relevant information about the origin of the data is conserved and communicated, enabling, but also shaping, their discovery and re-use. Researchers in adjacent fields rely on this data to further their work, but similarly act as collective reference points for the curation efforts. To provide the community with data that are trusted and usable, the expert work of biocurators serves to assure fit and quality (Müller & Naumann, 2003). Yet, curation capacities are not limitless.

Curators' workload-limits dictate the extent of the collection and the extent of the annotations performed (Rodriguez-Esteban, 2015). This is especially true for the resources discussed in this thesis: Specialized collections, providing, among other things, data on certain organisms, types of proteins, or genetic preconditions for cancer. Serving rather small communities, they are constrained by funding and rely, in most cases, on members of the community to perform a majority of the curation. This allows them to achieve detailed and relevant collections with small staff numbers. As experts in the field, these community curators come with a detailed understanding of the data they curate, ensuring that the knowledge represented is as correct as can be. However, as curating can be a tedious (Pinel et al., 2020, p. 189) and lengthy process (Chen et al., 2020, p. 98), the contribution might be limited. That means the curation process needs to be designed in a way that accommodates user needs and curators' motivation, amongst other things. A balancing act involving prioritization and careful consideration that is usually performed by those individuals who originally created and continuously maintain the resources. Next to users and curators, maintainers are an important group that often remains invisible (Linåker et al., 2024). They will become the focal point of this thesis and represent a majority of the people interviewed. Their work in designing and upkeeping biodata resources is shaped by prioritization, by different perceptions of the value of data in aggregated form and by connections to many different stakeholders (Tempini, 2017). The balancing of different values within the data curation, in interaction with biodata resources, and in their design, is the focus of this thesis. Connecting value with doing, my aim is to explore how data are shaped in the context of community-curated knowledge bases.

By doing so, I contribute to larger discourses on data aggregation, while more carefully extending the academic understanding of data work as multiple. Ranging from curation processes within data pipelines, to the doings of researchers, curators and users, and ultimately the infrastructural work required to enable them, data work will be placed in larger constellations, highlighting its valuative nature. Especially in extension of research on maintenance, my aim is to highlight the varied ways in which not just the data and their use-cases shape how biodata-resources are developed, worked on and maintained. In doing so, I will highlight the role of maintainers as central figures, who bring together data and people in knowledgeable, tireless and care-full ways.

To do so, I will first explore different kinds of data work, examining how they are conceptualized in the literature and how they are structured. In relation to the designing and upkeep of biodata

resources, I will further engage with notions of maintenance, especially software maintenance, before turning to valuation studies to build my conceptual foundation. This will provide the theoretical foundation, before laying out the methodical considerations and analytical approaches. The aim of this thesis is ultimately to identify the many ways in which data and data work are valued in and around biodata resources, highlighting how valuing is enacted in the practices performed. This approach extends the current state of literature on data work and maintenance for life-science data infrastructures. Choosing a valuation studies approach allows me to describe more than just data as situational, similarly exploring practices and infrastructures as enactments of value. In the data pipeline, in interaction with the data pipelines and in their design, various ways of valuing form complex connections. They shape and are shaped by many different practices, which are at times stabilized and at times shifting. These aspects will be explored in two empirical chapters. The first acts as an extension of the theoretical concepts, identifying registers of valuing, meaning shared ways to talk and think about data and data work. The second empirical chapter finally turns to the practices, systematizing them while drawing connections to the registers of valuing identified. Towards the end of this thesis, I want to focus my discussion on maintainers as central figures who manage these complex relations of data and value. Ultimately, I will place their engagement and the resources they provide back into the literature discussed, hopefully achieving my goal to extend the scope of how data work and maintenance are seen in the context of biodata resources.

2. The Social Science of Biocuration

I am entering a vivid debate. This thesis is placed in a time of high output methodologies, increasing open science awareness, and a public debate layered with big data and AI frenzy. Research data, biological research data, and their distribution, travel, and flow, have been discussed by scholars in Science and Technology Studies (STS) and adjacent fields over at least the last two decades (among them Bowker & Star, 1999; Kitchin, 2014; Leonelli, 2016). These are critical contributions to the understanding of data aggregation and use at large, but especially in relation to data work and maintenance. In this chapter, I aim to bring together their contributions, highlight important aspects for my analysis, while identifying gaps and missing links. Starting with explorations of data aggregation more generally and scholarship on pre-digital curation within the life sciences, will build a foundation for the current discourse and my own research. Introducing definitions of data and perspectives on data work, I then hope to portray the resources discussed in this thesis as socio-technical processes. Here, the multiple facets of that work will become visible. To further enrich this discussion, an additional body of literature will be introduced in the last section of this chapter. While data work is already often seen through the lens of care, this second group of authors highlights care-full doings of maintainers, especially in relation to the technical infrastructures that enable curation. Seeing data work and maintenance as entangled but multiple will enable a nuanced view into the doings and valuations of those working in, around and on biodata resources.

2.1 Two contexts

Before entering debates on data, data work and maintenance in their full breadth, I want to start by situating the work described throughout this thesis in two larger contexts. The first is the relatively new academic field Critical Data Studies, examining big data and the use of data aggregates. Here, scholars critically engage with the underlying assumptions and blind spots of data aggregation, at times even questioning the need for large-scale aggregation more generally. The second is the history of curation itself within the life sciences. Just as data are not new to the life sciences, neither are curation practices a novel phenomenon. By exploring a historical case of curation, the following data-focused discussion is related to pre-digital or at least less digital practices of similar kinds. The two perspectives serve as a backdrop for the more targeted discussion to follow.

2.1.1 Critical Data Studies

First named by Dalton and Thatcher (2014), Critical Data Studies explore questions of representation and power in relation to data, particularly in the context of big data. This scholarship presents itself as critical to the increased reliance on data as a representation of society, nature and the world at large (boyd & Crawford, 2012). Additionally, authors like Kitchin and Lauriault (2018) put into question the representative power of data, of large data aggregates, especially in the context of big data collection through online platforms. This data collection works differently in a few ways compared to the academic aggregation at the center of this thesis; however, the larger observations collected and concerns voiced by scholars in this field remain relevant to the questions asked in this thesis. The definitions of data and data work explored in the following sections will neatly fit in this larger context, though many predate the term Critical Data Studies. My aim is to make connections between scholarship on academic data and critical explorations of data aggregation more generally, to identify potential parallels.

The use of data for knowledge creation is not at all new, yet boyd and Crawford nonetheless identify a historical shift in the way that data are used (2012, p. 665). As methods change from small-scale analysis to data aggregation, systemization and synthesis, knowledge is produced on different grounds. More explicitly, the authors describe a shift from case-specific or sampled data to the collection of datasets aiming to capture each entity in question. This is especially relevant in relation to large aggregates of social and behavioural data collected by online platforms and other digital services (Gillespie, 2010, p. 355). Here, not only some users are represented in the dataset, but all users, as platforms follow their interactions at every point in time. This increasing engagement with regimes of data collection, integration and exploitation (further conceptualized by Zuboff, 2019) is the trigger for the emergence of this new field. While subsequent criticism is directed towards these exact types of aggregation and the companies performing them, critical data scholars raise larger questions, relevant to more than behavioural data.

As research shifts, at least in parts, to focus on aggregated data, the knowledge produced might shift as well. When aggregation is no longer a task embedded in everyday academic practice but separated

out into the data pipeline, scientific practice at large might shift (boyd & Crawford, 2012, pp. 665-666). The impact of this is described in a number of ways. For one, larger aggregates come with a higher level of standardization or if that is not the case, with a messier form, requiring more complex analysis and cleaning (boyd & Crawford, 2012, pp. 660-670). Each decision along the process further shapes and changes the aggregate, and therefore, the knowledge created. Without being aware of the data origins, users of aggregated data rely on their seemingly objective qualities, while previous decisions in terms of collection filtering, integration and format remain unconsidered. But the impact of these origins is not the only challenge raised by critical data scholars. Noting the sweeping nature of large-scale aggregation foundational to big data efforts, they question the notion of quantity over quality in general. Kitchin and Lauriault (2015) argue instead for an embedded collection that is targeted at local problems. Such collections, like the resources explored in this thesis, are more closely aligned with their communities – those are, the communities represented in the data and the communities using them.

Aiming to answer questions of representation, boyd & Crawford closely examine the relation between reality and the data describing it (2012, p. 667). They do so by asking more than just epistemic questions of representation and focusing instead on the practices of data collection and management, probing the power and ethics behind them. Here, data are followed as they move between contexts of data production and use, while the steps in between and even the underlying data formats remain opaque (Kitchin, 2014, p. 275). However, as platforms collect data invisible to users, which later shape interaction, Dalton and Thatcher ask who the beneficiaries of this collection and aggregation are (2014, p. 2). Central to answering this question seem to be control and access.

Data are not only collected by some for a more or less defined purpose, but also inaccessible to others who could benefit from their re-use just as much (Kitchin & Lauriault, 2018, p. 7). While the resources explored in this thesis all subscribe to ideals of open science and freely share their data while communicating their processes, not all research data infrastructures might. Even beyond access, questions of power and control still remain. The exploration of data processing of different types is central to these questions. As Kitchin and Lauriault (2018, p. 8) call for the tracing and unpacking of assemblages of data production, the biodata resource might be an ideal case to do exactly that. Their openness positions them far away from the platforms often critiqued (among them Facebook, Google and the like), yet it affords visibility to the complexity of data processing required for the re-use of any aggregated data. Here, too, different stakeholders are involved, shaping what data are aggregated, how they are curated and in what format they are subsequently packaged. By systematizing different positions, including curators and maintainers, as explored below, an additional level of granularity is added, highlighting at which points data movement is shaped and where it is put into action. In doing so, I hope to extend the understanding of data aggregation at large, providing a case study for opening up the data infrastructures and their underlying social processes. However, these processes do not necessarily need to be new. While the importance of large-scale aggregation is certainly reaching a

new extent, some of the underlying practices can be traced back to non-digital collections that previously shaped research in the life sciences.

2.1.2 Curation in the life sciences

Throughout the discourses on research data aggregates, many scholars point out that data are in no way a new addition to the scientific work in biology (e.g., Leonelli, 2016, p. 1). It is important to include at this point that, similarly, centralized efforts to support and standardize biological research are not new. Before researchers could access annotated genome sequences through online knowledge bases, another type of organization supported their work. To enable standardized experiments on organisms with known properties, stock centers played a vital role. These organizations housed large quantities of distinct species, all categorized and meticulously kept alive. In her 2019 study, Bangham explores the history of such stock centers, which provided *Drosophila*, also known as fruit flies, to labs that used them for genomic research. These centers are quite different in their output from the resources discussed in this thesis, providing small quantities of flies to labs on request instead of large quantities of data through always available interfaces. Nonetheless, the work associated with the one is a good starting point to explore the other.

Bangham (2019) describes the work at these stock centers as continuous yet flexible, as stock keepers did what automated systems apparently were unable to do: that is, keeping flies alive and, most importantly, also well. But working at a stock center was not just about keeping flies healthy. As new experiments required ever-new kinds of flies with new and different mutations, these collections needed to adapt their stock. Balancing capacity and space, such decisions required judging each species' relevance – adding new ones and discarding those no longer relevant. Once selected for integration, new species needed to be described in ways that specified their features while placing them into a larger overview of the stock available. In the wake of an increasing internationalization, these ordering systems developed beyond single stock centers, with some ultimately sharing their terminologies to make it easier to select flies and communicate experiments.

In her history of these centers, Bagham (2019) summarizes the closure of a number of such centers, respectively tied to one aspect of the work. The quality of the flies provided, the relevance of the stock maintained, and the way flies were described in more or less commonly understandable ways all lead to defunding and ultimately closures or consolidations of collections. The work with flies, the work on the description of flies and the work on the stock at large were all relevant to the upkeeping of such central institutions. Such types of work, in new contexts, will become relevant aspects in this thesis. Similarly, the different ways to make sense of this work, as providing healthy flies, good descriptions, and high relevance, foreshadow the different ways in which stakeholders can value what happens in data aggregation efforts. Without exploring further parallels, regarding, for example, funding struggles and recognition for the work of those stock keepers, my aim is to highlight that the work described in this thesis can not only be seen in the context of data work and digital maintenance. Flies have turned into data, but the work required shares a number of characteristics. Keeping in mind

the flies enables a connection from the history of the life sciences to the current state of biocuration. Yet, after pointing out these similarities, there are some differences to explore.

2.2 Data

While perceptions of fruit flies seem rather universal, the limits of what are and what are not "data" seem rather contested. Throughout the literature on data, their definition is often pointed out as a challenge. In this section, I want to explore the reasons for and solutions to this challenge and the role that distance plays in the perception and usefulness of data that travel. This will help further conceptualize some of the concerns raised by critical data scholars above and provide a solid foundation for engaging with different types of data work further below.

Scholars like Borgman (2012) or Leonelli (2016) dedicate entire sections of papers and chapters of books to the definition of data. At the center of their efforts to grasp the term is the realization that the word "data" is not an inherent property of representative objects; instead, "data may exist only in the eye of the beholder: The recognition that an observation, artifact, or record constitutes data is itself a scholarly act" (Borgman, 2012, p. 1061). It is understandable, therefore, that Leonelli defines the term in relation to the field of biology, or more precisely, in relation to the biologists that occupy it. Here, data are seen as "preservations of given phenomena" that are related to "the claims for which data can be seen as evidence" (Leonelli, 2016, p. 70). Such representations can, of course, come in many different forms, including representations of the works created through "observational techniques, registration and measurement devices, and the recalling, modification, and standardization of objects of inquiry" (Leonelli, 2016, p. 72). Such data are tied to their experimental setting, or when not experimentally captured, to the "instruments […], as well as the interest and position of the observer" (Leonelli, 2016, p. 77). Between the representation and the claim for which they are used, data are tied to two contexts: the context of their creation and the context of their interpretation. The movements between the one and the other…

> *do not merely affect the interpretation of data, they determine what counts as data, and for whom. As data travel, their producers, curators, and users keep making decisions about what constitutes data in relation to the changing circumstances and aims.*
>
> *(Leonelli, 2016, p. 77)*

Not surprisingly, these definitions of data may vary depending on the distance that data travel. Such a distance is not necessarily just a matter of physical space, but can come in a number of forms. When the producer and user are one person, the context might vary only in relation to time and space, as the location of the lab and desk, and the time difference between lab days and desk days, require the data to move. Including this case, Borgman & Groth identify six axes on which data can move great distances: "domain, methods, collaboration, curation, purposes, and time and temporality" (2025, p. 6). The most complex of these terms is the domain, describing the entire epistemological and socio-technical culture of the originating and data-using research community. This axis represents the complexity of distance, beyond the other axis identified, as it highlights the many subtle ways data

can be perceived. These cultural differences mean that scholars across fields may view data quite differently: “Data and information can be mutually understandable within a user group, but may not be interpretable outside that group because of unfamiliarity of specific language, scientific symbols, and data formatting structure” (Huang, 2015, p. 19). Such inabilities to interpret and therefore use data are what Edwards et al. call “data frictions” (2011, p. 669), the results of changing contexts through different ways of moving. Returning to the axes of distance, this implies differences in methods, and the purpose for collecting and using data must be translated, shaping how data can be utilized across distance (Borgman & Groth, 2025, pp. 9-10, 15-16). But the travel itself also plays a role. In the context of collaborative distance, Borgman & Groth describe the connection that the data producer and the data user have (Borgman & Groth, 2025, pp. 10-11). As a bridge between cultures, the two can help overcome distance if they are themselves close, but might create more friction if they are not. Similarly, curators, doing translational work, can bridge the gap. Their role will be described at length in this chapter, but in regard to distance, the authors highlight the importance of their position, as it can be closer to the creation or closer to the use. The impact of this position of the data curator is what Edwards et al. (2011, p. 673) describe as metadata frictions. As translations are done with one or the other scientific culture in mind, they are once again more or less attuned to the creation and use of the data. Data, therefore, become tied to a third place where data work happens.

2.3 Data work

The concept of data work is not in itself new, but in the wake of online platforms and so-called click work, it has gained additional layers over the last years. Originating in scholarship on the organizational system of standardization, data work of different kinds can be traced to Bowker and Star’s *Sorting Things Out* (1999). In talking about knowledge in international organizations, the two discuss the struggle of distance mentioned above and highlight the work necessary to bridge those gaps. However, their exploration is still very much focused on the epistemology and less on the work itself. This focus is set some years later by Karasti et al. (2010), when discussing the work necessary to develop information infrastructures and move data through them. While only mentioning “metadata work” explicitly, they highlight associated practices of data work throughout their paper. Tracing information infrastructure in the context of a long-term research endeavor, with short-term projects, they describe how data work is navigating different temporalities and is working across sites (Karasti et al., 2010, pp. 395-396). Here, those in charge of managing the data are asked to accommodate many different needs, shaping not just their work but the infrastructure they are working on (Karasti et al., 2010, p. 397). Before exploring data work in such an infrastructural sense, I want to turn away from academic data for a moment to look at data work from a different perspective. Here again, Critical Data Studies and adjacent work provide a great introduction into the processing of data, highlighting the more than neutral doings within aggregation efforts.

Outside of academic knowledge infrastructures, online platforms in the late 2000s claimed to move information neutrally from one place to another (Gillespie, 2010, p. 358). Yet to keep their “neutral” online spaces clear, they required more than automated systems to decide what information was

"okay" to transport and what information should not be platformed. This work, often outsourced to people outside of platform companies, is part of a larger field of supposed automations carried out by humans through digital labour. Described by Irani (2019), this work of filtering content, recognizing images, and accomplishing small tasks is placed in a larger analysis of the labour required for a neutral flow of information. "These cultural data workers are central to the political economy of computing, from the free labor of AOL chat-room moderators to the organized but invisible labor of paid content moderators" (Irani, 2019, p. 27). Such tasks can be linked to the knowledge data work mentioned before, as digital laborers also help the system to place a piece of content in a cultural context, when, for example, tagging it as appropriate or not appropriate. However, the notable aspect for Irani is not the epistemological work accomplished by click workers; instead, she focuses explicitly on the working conditions and the structures under which such labor is performed (2019, p. 29). In a similar vein, Gregg and Andrijasevic (2019) analyze how digital work is coinciding with "an active celebration of worker flexibility" (Gregg & Andrijasevic, 2019, p. 3). By highlighting these working conditions, the authors ask not only what kind of work is performed, but also under which circumstances, and by whom. As part of this analysis, Gregg and Andrijasevic differentiate between the mostly male employees of tech companies who build data work platforms and devise tasks, and those who perform them, often women and members of marginalized communities (Gregg & Andrijasevic, 2019, p. 3). This division of labor between structural work and data work will become relevant for the further development of my research. Yet for now, I want to explore the notion of invisibility.

One of the most prominent platforms for the management of digital labour is called Amazon Mechanical Turk, a name referencing an automatic chess-playing machine that is secretly run by a human sitting inside it. In the case of the platform, this invisibility is a feature. "With workers hidden in the technology, programmers can treat workers like bits of code and continue to think of themselves as builders, not managers" (Irani, 2019, p. 34). But even in cases where invisibility is not the goal, data work is easily hidden from those using a platform or dataset. Returning to academic data, this becomes a challenge for the curators of research data, as their work remains without credit, and their efforts and struggles remain invisible (Davies, 2025, p. 5). As crucial parts of data infrastructures, they help to enable movement across contexts for scientists who, at times, are unaware of their existence. As Davies and Holmer (2024) note, this underrecognition is not just a matter of emotional frustration but, in turn, affects the visibility of the epistemological work they perform as well. As biocruators order and make movable research data, their contributions to terminology and ordering standards remain underrecognized (Davies & Holmer, 2024, p. 693). Once again, this invisibility comes with struggles for funding and recognition in an impact-driven academic system. Returning to the notion of distance, this means that those in the third context – the place of translation between creation and use – become invisible and constantly threatened contributors to the shape of the data, despite their crucial role in making data moveable across distance in the first place.

Just as stock keepers ensured that flies were not only alive but also healthy and relevant, data work extends beyond the digital relocation of data between contexts. To provide "good", "usable", even "pristine" (Plantin, 2018) data to researchers, curators interpret the large amounts of messy data presented to them. Here as well, the interpretive efforts are work described by scholars through the language of care, as curators ensure that distance is bridged as well as possible (Davies & Holmer, 2024, p. 693). *Care* as a concept is used in this context to make visible the mundane and underrecognized, yet crucial work of curation. This work takes the need to worry, or even think, about frictionless data movement from those "doing research," while providing the foundation for their very accomplishments (Davies & Holmer, 2024, p. 695). This description of data work as care work highlights the more than mechanical nature of fitting everything from experimental data and genome sequences, to published and peer-reviewed research papers, into a structure best suited to them, their scientific origin and their potential use.

Keeping in mind the invisibility and care with which this data work is conducted, it is important to specify what this work entails. Explored, among others, by Leonelli throughout the years (e.g., Leonelli, 2009, 2014a, 2016), data work comes in a number of forms and is positioned in varied places. Between creation and use, data need to be de-contextualized and re-contextualized before being ready for re-use (2014b). The challenge of this endeavor, however, is to convey the key aspects of data curation while balancing between the differences of the two contexts and what can be assumed as shared understandings (Leonelli, 2014a, p. 10). The aim of these efforts is not just to make data **R**eusable but also to enable better **F**indability, **A**ccessibility and **I**nteroperability, or to make them "FAIR" (Wilkinson et al., 2016). Throughout the years, the need to balance information that is commonly understandable, consistently defined, and movable between contexts has led to the development of shared ways of describing data and data creation, but developing such standards is itself challenging. Describing the development of such systems, Leonelli points out the need to accommodate relevant information in a changing research environment, while keeping internal consistency and stability over a longer period of time (Leonelli, 2009, p. 6).

To make a number of such ordering systems interoperable, coordination is needed between curators, further expanding the potential reach of data. Through collaborative work, shared ontologies can be developed to better describe the data, while specialized ontologies are optimized for integration. As a "map of reality used by scientists to coordinate and share resources" (Leonelli, 2009, p. 7), ontologies build the structural foundation for much of the work described in this thesis. Yet, their development is itself a huge undertaking, requiring those involved to make explicit what is meant by terms that are supposed to travel across contexts (Rubin et al., 2006, pp. 187-188). As contexts and research develop constantly, these ontologies too have to change to ensure a smooth translation (Leonelli et al., 2011). This becomes especially visible when description systems and data do not match, limiting the scope of what can be transported. As described by Plantin (2018), such cases create a situation where movable data is only useful within the scope considered in the development of the ontology. Once standards are set, the curation might entail leaving out relevant aspects of the data from their

description, despite potential relevance for downstream uses (Plantin, 2018, p. 60). In this case, curators themselves are aware of the limits of their descriptions, but are bound by the vocabulary available. As ontologies are best developed before the curation to assure consistency for data created across time, the ontology work and curation become two important aspects of data work that need to be considered as entangled yet temporarily separate (Leonelli et al., 2011, p. 7).

Both curation and the building of ontologies are not necessarily tied to a specific epistemic place, or even to centralized efforts. In the case of the decentralized long-term ecological study described by Karasti et al. (2010), data packaging and description are tied to a project and its short-term sub-projects. Here, already, questions of longevity and fit to short-term needs require substantial work in coordinating the formats of data available and their contexts of creation with potential uses and long-term, standardized storage (Karasti et al., 2010, p. 404). What complicates this endeavor further is that data curation and the development of metadata standards need to be implemented into a technical system that facilitates the movement of data between sites.

The development of technical systems moving data between contexts is the last element of data work I want to highlight in this section. While curation and the development of standards prepare data to be movable, they need to be closely integrated into digital infrastructures, facilitating the movement itself (Karasti et al., 2010, p. 405). Outside of local contexts, these digital infrastructures are what make data travel. Enabling the sharing ideals of open science, databases, repositories, and knowledge bases bring together the data created in multiple sites, facilitating curation and distribution for re-use (Leonelli, 2016, p. 14). Referencing research on model organisms, Leonelli and Ankeny note elsewhere that databases "made it possible to dramatically increase the quantity of information on model organisms that can be stored and integrated, as well as the number of researchers with access to such information" (Leonelli & Ankeny, 2012, p. 34). The development of such resources, however, is more than the implementation of ontologies into a database that accepts certain file types. As Borgman and Bourne (2022) remark, creating and making viable digital infrastructures can require extensive effort. This work is ultimately tied again to the contexts of data creation and data use, as "many individuals are engaged in data-handling" (Borgman & Bourne, 2022, p. 4), all of which need to be considered in the development process. Their accounts highlight the need to not just fit data standards to the needs of contributors and users, but to align the digital infrastructures themselves. Reflecting on such processes of building a resource, Murillo (2023) stresses the importance of making standards and their implementation fit the community that is served. Here, "common" and "public" are not seen as the same, highlighting once again how resources and their terminologies can bridge between people in some research communities but not across infinite distances. This work of connecting communities on a technical level is once again a question of care-full engagement. Ranging beyond epistemological work, curators engaged in building digital infrastructures need to consider a wide range of things, including how to motivate contributors and how to allow for crediting of both the resource and those originally collecting the data, all while securing funding and stakeholder support (Borgman & Bourne, 2022; Levin & Leonelli, 2017). Still, the maintainers' work

is often described in the same breath as the data work described before. Understandably, many scholars (e.g., Burge et al., 2012; Davies & Holmer, 2024; Leonelli, 2016) describe the collection, handling, collective harmonization and hosting of research data under the umbrella of "curation". This description of the work of "database curators" (e.g., Leonelli, 2016, p. 50) as a singular entity is understandable, as the different types of work explored in this section are often integrated into the responsibility of individual people or small groups that make data move.

Going forward, I argue that the distinction between different practices in managing research data can be rather fruitful. In the case of click workers mentioned earlier (in reference to Gregg & Andrijasevic, 2019), the distinction between programmers and data laborers is a matter of visibility and working conditions. Similarly, in Plantin's (2018) account of curatorial data work, the distinction between curators and managers is mostly in relation to visibility. Here, curators are invisible to the users, but their actions are very visible to their managers overseeing the curation. This second study, however, comes with another important aspect. As curators find better ways of dealing with data than is envisioned by their curation manuals, they hit the limit of their agency within the process. To provide a "pristine" dataset, managers limit the scope of possible curation actions, opting for consistency and not the best representation of each curated object (Plantin, 2018, p. 11). While curating is necessary to make the data move, so are the more managerial tasks of designing and upholding the data pipelines in which this curation happens. This distinction of positions and the varying requirements they face will be a crucial distinction going forward. While doing so under the umbrella term of "curation," Leonelli (2016) nonetheless describes in great detail how data are aggregated, formatted, annotated and circulated, while highlighting these managerial tasks. In her book *Data-Centric Biology,* she lays out how decisions about the flow and reshaping of data come to be under the involvement of varied stakeholders. Rather than seeing "the data journey" as neutral, Leonelli (2016, p. 178) emphasizes it as a site where scientific reasoning and institutional constraints intersect. Similar to this thesis, her approach explores how data are seen throughout the process, describing the value of data through academic, political, financial and affective lenses (Leonelli, 2016, pp. 63-64). In my exploration, I want to extend her work in particular and the work collected so far in this chapter in general. This involves differentiating between various types of practices and highlighting where and how decisions regarding the movement of data are made. By identifying different positions and their ways of seeing data and data work as valuable, I hope to contribute to the visibility and to a structured understanding of data work in biodata resources. To position my work in relation to yet another body of literature, the next section will explore notions of maintenance to better grasp the work of those who ensure longevity for digital research data infrastructures.

2.4 Maintenance

Not all data work is aimed at longevity. When individual researchers aggregate and prepare data for their own use, this work is aimed at a certain, possibly one-time use. What is described as "mundane data work" by Ribeiro et al. (2023) is a tedious engagement with data that needs to be moved from its creation process to its re-use (Pinel et al., 2020, p. 189). Central to the collective data aggregation

efforts at the center of this thesis is to make such work beneficial beyond one use case. By making data movable and sharing them, they can serve a number of causes, and by standardizing the ontologies used for this description, these data can travel even further and potentially be integrated with others. Nonetheless, in this process, the mundane work of making data movable is still crucial. As described above, it might become invisible while simultaneously increasing in complexity. But invisibility in union with internal complexity is not just a feature of biocruation. Infrastructure in general is often described in relation to its decreasing visibility (Star, 1999, p. 381). Describing the aggregation, integration and distribution of research data as such an infrastructure is not uncommon, both as an explicit framing (see Baker & Yarmey, 2009; Linåker et al., 2024; Tempini, 2017) or in less focused ways throughout the literature already explored above. Between the "mundane" and what Tempini (2021) calls "innovative data practices", my aim is to explore the nature of maintenance in general and in the case of research data infrastructure specifically. In this section, I focus on this view of data work, specifically in relation to the efforts required to ensure longevity.

In an era of innovation and the celebration of innovators, two essays from Russel and Vinsel (2016, 2018) are aimed at shifting the focus from innovation in technology to the maintenance of technologies. Especially in their later publication "After Innovation, Turn to Maintenance," they highlight the discursive absence of "mundane labor" or "normal science" (Russell & Vinsel, 2018, p. 7) work, which is omitted from historical accounts of technology. They attribute this invisibility to two potential reasons:

> *In some cases, maintenance is absent from historical accounts because historians have ignored the maintainers. But in other cases, societies simply did not devote resources to maintenance, even when they understood it was in their own interests to do so.*
>
> *(Russell & Vinsel, 2018, p. 9)*

But invisibility is not just a question of the outside perspective and societal neglect; it is similarly tied to the systematic structuring of maintenance work itself. Denis and Pontille (2017) explore different societal regimes of maintenance and highlight how different perceptions of technology influence what is and is not widely perceived. They, too, connect invisibility with a focus on innovation and "things that are supposed to stay stainless and flawless" (Denis & Pontille, 2017, p. 4), but here, the invisibility of maintenance is not just a result of lacking attention but of an active process aiming to keep up the perceptions of perfection and immutability. As a result, there are two groups of people: those for whom the infrastructure is invisible until breakdown (as conceptualized by Star, 1999, p. 382) and those who know its parts and potential sources of error, performing maintenance whenever a need is anticipated (Denis & Pontille, 2017, p. 6). Similarly, data work remains invisible to users interacting with digital infrastructures that provide structured and aggregated data in a movable format, while biocurators perform the work necessary in collaboration with peers, all of whom see what users might not (Quaglia et al., 2022). Denis and Pontille, however, offer a second possible approach to maintenance. Where "stainless" innovation is not the focus, maintenance can become the task of everyone, bringing with it the visibility of upkeep and change (Denis & Pontille, 2017, p. 6).

In this case, the object maintained is rendered as requiring work and as malleable to all. Seeing maintenance through these two lenses will become important throughout the remainder of this thesis. While representing two extremes – completely shared responsibility and completely hidden maintenance – this perspective enables a description of data work that is, in part, invisible and, in part, a community effort. Coming back to the differentiation between curator and manager, this allows a more nuanced view of different kinds of data work that are visible to some and invisible to others. This is especially the case when curation, but not maintenance, is performed by the user community.

Before fully connecting maintenance to the data work discussed above, there is a need to explore it in the context of digital infrastructures, or more broadly, software, first. While physical objects are tied to notions of persistence, due to their materiality, software is more fluid (Ensmenger, 2014). Discussing maintenance for software, Ensmenger points out some special features of software that need to be considered. The underlying code that structures and makes a program does not change over time. It can be replicated, moved around or left alone in a way that physical objects can not. Nonetheless, code can break, as edge cases and environmental changes make its functions fail (Ensmenger, 2014, p. 3). The result is a perspective on software as a relational system, allowing for the consideration of contextual changes and, more importantly, of changes to the software itself. This is important as the "finished" state of software is described by Ensmenger as an arbitrary status (2014, p. 5). Changes and fixes can be made continuously. As contexts and user needs change, maintained software enters a constant state of alteration, fixing previous errors and adapting to new contexts, while advancing in its feature set. Maintenance of software is, therefore, not a constant state of repair (as described for infrastructure generally by Graham & Thrift, 2007), but of evolutionary, ongoing improvement with increasing complexity under the threat of decay (Lehman et al., 1997). Decay, in this case, is related to decreasing usefulness in the wake of changing needs. As software can always change, maintenance then includes the adaptation of the software to keep it and its users' interests alive.

Due to the temporal fluidity of the building and maintenance of software, maintainers are faced with increasing complexity (Ojala & Leavitt Cohn, 2024, p. 166). Both in systems not initially nor properly designed to withstand change and in systems with a long evolution, software maintenance has to be seen as knowledge work, designing and changing parts of the system in ways that fit the previous development (Ojala & Leavitt Cohn, 2024, p. 167). This work itself is shared between humans and non-humans, including guidelines, documentation, and versioning tools. Design and maintenance are then a question of cooperation with current and future peers, to ensure longevity and structured continuous development. Ojala and Leavitt point out the central role of builders and maintainers in this work, who become key figures, tasked with collecting and holding key knowledge of the system that is crucial for sensible further development (2024, p. 179).

Collecting this literature on software maintenance has two uses for this thesis. For one, biodata resources are themselves software that needs to be maintained. However, there is a broader connection that can be drawn. While ontologies and data collections are not strictly speaking software, they share

a lot of features, including their constantly evolving state, and the shifting needs and contexts they need to serve. Maintained for longevity, yet faced with short-term challenges, both data work and software maintenance are a process of balancing, with a strong need to document changes, coordinate with peers, and integrate into a larger system. This is where mundane maintenance turns into "innovative data work" (Tempini, 2021), seemingly reversing the "turn to maintenance" (Russell & Vinsel, 2018). However, Tempini describes a practice that is divided into two. Laying out the work on a medical database, he first defines data work in the sense of curation, highlighting its standardized nature within already established systems. Yet faced with funding struggles and calls for further development, a second type of data work emerges, aimed at further developing the resource to ensure longevity under the pressure of changing contexts (Tempini, 2021, p. 83). This innovation differs from the one critiqued by Russel and Vinsel (2018) in that its aim is not the replacement of the existing resource, but rather the adaptation to changing needs under the threat of defunding. The result is a system of exploration and explanation, where maintenance is a function of innovating outside of existing functionality, while integrating what has become established into a sustainable system (Tempini, 2021, p. 90). Such structured, constant evolution is also central to the development of ontologies. Summarized by Leonelli and peers (2011), the maintainers of the Gene Ontology rely on its underlying structure that enables constant adaptation to new needs and even large-scale changes. Here too, technical knowledge is required, next to extensive familiarity with the ontology in its current and past states (Leonelli et al., 2011, p. 6). Cooperation between peers and documentation for future maintainers is central to this process to…

> *produce a faithful representation of knowledge domains as they keep developing, which requires the translation of general guidelines into specific representations of reality and an understanding of how scientific knowledge is produced and constantly updated.*
>
> *(Leonelli et al., 2011, p. 7)*

Maintainers in the case of biocuration, therefore, deal with knowledge on two fronts. For one, they navigate the internal knowledge to enable further development and upkeep, and they manage the scientific knowledge they are trusted with, shaping structures and vocabularies of knowledge representation. Perceived or not, their role comes with a responsibility to make data movable in a way that makes them applicable to other places, but not without losing their underlying connection to reality. This "epistemic responsibility" (Corlett, 2008) entails accountability to those utilizing the aggregated knowledge, layered with notions of trust and liability. Yet, maintaining a database is not just a function of representing knowledge.

As mentioned throughout this chapter, maintenance for biodata resources is tied to a large number of different needs and contextual requirements. Following the ongoing development of a data collection platform for patients, Tempini (2017) provides an overview of the different factors that maintainers need to navigate to create a "valuable" product. "Valuable" in this context is not a function of monetary value but is seen along the lines of many possible "dimensions of value" (Tempini, 2017, p. 2). This approach is similar to the concepts used in the following parts of this thesis, allowing for the

analysis of maintenance work through a number of lenses, each with its own definition of valuable data, data work and data infrastructures. Tracing the longer-term development of the resource, Tempini identifies four such dimensions that influence and are the outcome of ongoing maintenance: Data collected from patients can be valuable to businesses developing drugs, can generate scientific value as source material for research, have the capacity to build community through the exchange of experiences and might help individual patients tracking their own long-term health development (Tempini, 2017, p. 10). What becomes visible through this approach is the influence of each of these dimensions on the development process. Each representing a different group of stakeholders, these ways of seeing data shape the development, both short-term and long-term. Connecting social to technical aspects of software maintenance, ongoing development is not just a function of these ideals of the platform; instead, technical challenges need to be negotiated with long-term epistemic representation and the value-laden social expectations, making infrastructural changes both necessary and problematic (Tempini, 2017, p. 21).

Outside of biological research, Cardoso Llach (2019) describes the development of shared technical resources and their metadata structures as the result of conflict. He highlights how different stakeholders of a shared project influence every aspect of the collective data handling. This ranges from decisions on what data are stored, and the way they are stored, all the way to questions of visualisation and subsequent use. The entire pipeline and everything that moves through it is the result of technical, but especially social, negotiation, heated debate and different levels of agency in a hierarchically structured group of collaborators (Cardoso Llach, 2019, p. 457).

Negotiating the different ways that data and data work are valuable is a central aspect of the maintenance work. Returning to Borgman and Groth's (2025) description of the distance between data origin and the context of data use, this means that maintainers are not just tasked to make data movable through their resources. As data pipelines through which data travel, these resources represent a crucial context for the reshaping of data themselves. This translation enables re-use across great distances, yet is similarly the result of many different negotiated needs, ideals, and contextual requirements. Not just the preparation of data for travel is a question of care-full doing; so are the building, coordination, and maintenance of the ordering systems, as well as the development and upkeep of the technical infrastructures required (Baker & Yarmey, 2009, pp. 13-14). Even more, the ways that all of these activities can be seen as valuable reach far beyond epistemic questions. Peers, users, funders, institutional contexts, and more all shape the requirements for data work of any kind, while not much more than the output is ever widely perceived (Borgman & Bourne, 2022).

Maintainers take responsibility for data as well as outcomes for any of the mentioned groups, to ensure sustainability and long-term usefulness. However, driven by ideals of openness and data sharing, this can come into conflict with more personal needs as described by Linåker et al. (2024). Centering the role of maintainers and contributors, the authors present work on data infrastructures as emotional and care labor, further problematizing the invisibility described by Davies & Holmer (2024). Here, importantly, Linåker et al. distinguish between different positions in their description of

emotional needs and potential strategies to sustain engagement. Maintenance of shared resources is described as pressuring and burdensome, recognizing the work and, more importantly, the energy that maintainers put in (Linåker et al., 2024, p. 5). Here, too, funding struggles and heightened expectations from stakeholders are identified as sources of stress. Making matters worse, maintainers' responsibilities also involve the management of workload and capacity for outside contributors. These peripheral figures similarly go out of their way to work on shared resources, often taking over more structured tasks to keep running and further develop a resource. Their engagement, too, is driven and limited by personal needs and capacities (Linåker et al., 2024, p. 6). This group, neither solely using nor strictly maintaining the resource, closely represents the community curators who engage with the resources discussed in this thesis. Their work is structured by the work of maintainers, but their capacities and emotional attachment to the resource might be limited. While paid curators, as in the study of Plantin (2018), are incentivized by their employment to sustain even emotionally unfulfilling work, volunteers from the community and their emotional needs have to be considered more carefully in developing things like curation guidelines. These more-than-data-focused elements of maintenance will become a central element of this thesis. The distinction between the emotional stress of maintainers and the emotional capacity of curators will add an additional layer to consider in the analysis, and especially in the discussion later on.

Maintenance, as explored through the scholarship in this section, is therefore more than just the repair of the technical object. As software and other infrastructures can and need to constantly be reshaped, maintainers are tasked with upkeep as well as development. Their responsibility goes beyond the transportation and translation of data between one context and another. Resources and shared ontologies similarly require their attention, all while stakeholders and shifting contexts present ever-new requirements. This notion of maintenance goes beyond the mundane data work required in preparing data for movement. Instead, resources need to be in a constant state of update and innovation internally to present and enact longevity in their output and towards the contexts they connect. While similarly a care-full process, this rather structural data work can be differentiated from other data work, enabling a more nuanced view of the processes that shape and are shaped by biodata resources. This differentiation of practices enables a further distinction between positions, in and around the data pipeline, separating curators and outside contributors from maintainers and the stakeholders they engage with.

As described by Plantin (2021), Tempini (2017) and Linåker et al. (2024), the ways that curators, maintainers and other stakeholders make sense of data and data work hugely influence the way that work and infrastructures come to be. By seeing practices in relation to values that are shaped by considerations beyond epistemics or economics, this distinction of positions becomes even more fruitful. This focus on the practices and their connected values is in line with D'Ignazio and Klein's (2020) call to make visible the labour shaping data, problematizing what is described in the previous sections as hidden work (Davies & Holmer, 2024) and digital labour (Irani, 2019). As part of their paper "Seven intersectional feminist principles for equitable and actionable COVID-19 data,"

D'Ignazio and Klein explore similar notions to the ones mentioned so far in this section, highlighting questions of power, objectivity and contextualness. Beyond that, they call to challenge archaic systems of classification and, most importantly for this thesis, to highlight the emotional and embodied nature of data practices (D'Ignazio & Klein, 2020, pp. 3-4).

By systematizing this analysis of values and practices in relation to each other, a better understanding of biodata resources will hopefully emerge. As the resources focused on in this thesis are, for the most part, not large repositories but community-embedded collections of knowledge, they act both as a representation of data work and as cases for digital infrastructures that are potentially more aligned with their surroundings. I therefore aim to position my research as a case study for examining data work more broadly, following the call of Critical Data Studies scholars to highlight the practices and values underlying data aggregation. At the same time, the community-driven nature of the resources will be highlighted, and their embeddedness into scientific communities might position them as a positive example for responsible and purpose-driven data work. Exploring this work exactly, I join scholars like Karasti et al. (2010) in differentiating between different types of data work and Plantin (2021) in the recognition of the varying positions involved. In doing so, I hope to further differentiate the many types of data work.

Following Leonelli (2016) and Tempini (2017), these practices will be explored through the lens of values, shaping perceptions of data and shaping practice and management alike. To focus value in this sense, the concept of care, as centered by (Zakharova & Jarke, 2024), will be expanded on briefly in the coming chapter, highlighting especially subjective and embodied material practices. Throughout, maintainers of constantly evolving digital infrastructures will be central figures, as their version of data work is crucial to navigating, synthesizing and negotiating the different values at play. Their engagement with cooperatively designed ontologies, outside funding structures, and, most importantly, with their respective research communities needs further systematization, which I hope to provide going forward. Care-full data work shapes more than just the data pipelines it constructs and upholds, but the data that runs through them and the aggregates they form. As such, maintenance work is central to the bridging of contexts in which data exist, are created, translated and used – again and again.

3. Value and Practice

Care is not the focus of this chapter, but care is where I want to begin. In building a theoretical foundation that will guide the rest of this thesis, care will be an underlying foundation, supporting my explorations of valuation and its entanglement with practices. In this first section, care will be the connection I draw to the literature mentioned above, providing a mode of seeing the world, its human and non-human inhabitants and their doings of any kind. In this chapter, care will allow me to build my understanding of value and valuation in a way that is conceptually solid and tied to its pragmatist-materialist roots in STS. As central features of the concept of care are similarly central to valuation, they require their introduction in union. Care-full doing and valuing through practice are intertwined.

Departing from care, the following sections of this chapter will embrace valuation as a central concept for understanding practices of any kind and for structuring the remainder of this thesis. By exploring its underlying assumptions, empirical affordances, and related concepts, all relevant facets of valuing for the following analysis will be assembled. As an interplay of evaluation and valorization, all practices are seen as contextually embedded and positional, opening up the possibility to look at them as multiple and changing. By introducing registers of valuing and valuation constellation as extended perspectives on valuing, I hope to achieve the goals set for this thesis. Valuation, yet especially the registers of valuing and valuation constellations, will shape the structure and analysis presented in the empirical chapters below.

3.1 Building on Care

Care is about doing – the "doing of," the seeking out, the acting with and maintaining of; the supporting and upholding. Care is an attentive practice (Mol et al., 2010, p. 10). It is an ongoing process, separate from universal notions of success and failure; an "activity that includes everything that we do to maintain, continue, and repair our 'world' so that we can live in it as well as possible" (Fisher & Tronto, 1990, p. 40). Care is not forward-thinking or productive per se, but can similarly be attentive, compassionate, and generous. It does not have to be limited to humans, and it does not need to be one-directional. Care involves all kinds of beings and bodies caring for and being cared for. It is situated, local, personal and embodied (Mol et al., 2010, p. 10).

In its original use, the term is associated with doings of "care work," describing practices of "caring for" and "catering to" (Mol, 2008). First described in the analysis of gendered work and placed especially in family or medical settings, where caring is more than just a set of predefined practices. But caring occurs in many more settings, and as such, its use in STS extends beyond care work (in part assembled in Mol et al., 2010). Looking at care-full practice enables a new understanding of doings of all kinds, ranging from the very concrete data work mentioned above to theoretical explorations of knowledge production itself. Just as data workers caringly conduct the invisible labor of aggregating data, so do scientists of all kinds carefully assemble their representations of reality (Puig de la Bellacasa, 2011). Throughout these practices, the term provides the same critical lens, focusing on embodied material interaction. Such care-full doings, however, are highly situated and personified. The questions "When?", "How?" and "For whom is cared?" make visible the subjective dimension that comes with care (Mol et al., 2010, p. 15).

Selectively, care is directed at something, for a certain reason, with a subjective intention. Attached to the practice is an ideal, a way to value. But this caring engagement comes with a counterpart, described by Martin et al. as the "dark side of care" (2015, p. 627). The absence of "caring for", the avoided engagement, and the refusal – both active and not – to strive towards something, is as much a part of this positionality. At the heart of care is the performance of a normative "good" and "bad," an attending to what is important, not just a doing of what is necessary. By describing care as a "selective mode of attention" (Martin et al., 2015, p. 627), it inherits elements of power, the pursuit of something

and exclusion of something else. More relevant here, this makes caring practice a function of the values attached to it. As every act of care implies a hierarchy of attention and prescribed worth, valuation becomes a central element worth exploring – at least in the context of this thesis and for the purposes of my engagement with biodata resources.

3.2 Looking at Value

Under the headline (and journal title) "Valuation Studies," scholars from STS and beyond have assembled and further developed the analytical lens of value in the last decade (as summarized by Asdal et al., 2024). In their exploration, they depart from previous notions of value that are mostly tied to economic value and processes of value creation and appraisal, aiming to redefine what value and valuation stand for. In their introduction to the first issue of "valuation studies," Helgesson and Muniesa (2013) reintroduce value as something that is prescribed, not inherently tied to an object, something or someone; as something that is multiple, changing from context to context and person to person (Helgesson & Muniesa, 2013, p. 7). For them and like-minded scholars, value is relational and, most importantly, enacted, meaning it is rooted in practice and continually performed (Dussauge et al., 2015, p. 8). This view on value traces back to theoretical considerations using Actor-Network Theory to explore economic processes and markets (Asdal et al., 2024, p. 80). As collective enactments, actors in these contexts create value through their practices.

The resulting theoretical body affords a number of relevant considerations that I want to explore in this section. As value is not inherent, but prescribed, evaluation becomes a matter of embodied and contextual assessment. Value is attached to something according to an individual, or through a possibly shared perspective. Evaluation in this instance is not just performed in an economic sense, but might come in a number of different forms (Vatin, 2013, p. 33). Value is not just ascribed; it is also created. Through targeted practices aimed at heightening any kind of value, the evaluation can change. This valorization is once again driven through subjective ideals of value, as their impact is targeted towards an improvement of a certain type (Vatin, 2013, p. 33). Both evaluation and valorization need to be seen as situational, embodied and rooted in practice, seeing and creating value in a certain way. Exploring the two terms, Vatin (2013) further entangles these doings. As both evaluation and valorization are rooted in subjective ways of seeing – one concerned with identifying value, the other with creating it – the two terms need to be considered as part of one practice. As value is ascribed through evaluation, the value itself is created. Similarly, the improvements in value are based on a perception of the current and the prospective value, meaning they are based on evaluations. Combining the two terms into "valuation" as a hybrid term, practices of any kind can be described in relation to their evaluative and valorizing effects (Vatin, 2013, p. 47).

As valuations are seen as relational, so are the caring or less caring practices tied to them. Following this logic, the doings of any kind are then the representation of underlying embodied, situational valuations. This combination of value and practice, each as relational, allows for an analysis of all sorts of activities with consideration to ideals of, emotions on and perceptions of value. Practices are

tied to value. This connection is crucial for the analysis ahead, as different relations between humans and non-humans are involved in data work, which can now be assessed in relation to their shared and not-so-shared ways of valuing, described under examination of their collective and individual practices. This can range from perceptions and performances of measurements or research papers as data, to ideals of what good data aggregates, good data work and good data infrastructures might be, should be and need to be enacted as. Disentangling these practices and values, however, can present a challenge, as many factors need to be considered, including the researcher's position, practices and valuations themselves.

3.2.1 Registers of Valuing

To reintroduce some structure into this exploration of valuation, Heuts & Mol (2013) aim to categorize different ways of seeing and attributing value that share an underlying logic. In their study of *good* tomatoes, shared *registers of valuation* emerge between different actors, situations and contexts. By their interviewees, tomatoes are perceived in relation to costs, handling, nostalgia, naturalness, and sensation (Heuts & Mol, 2013). However, each person's experience and descriptions of these terms vary slightly, leading to different practices. What is important in this case and in relation to registers of valuing in general is that a certain register is always in relation to the individual perception of the object. In a case where the tomato is destined to become ketchup, its "naturalness" is mostly connected with communicative aims, while the growing process is, in turn, highly engineered. In the case of a home gardener, this naturalness has different meanings. Registers of valuing are, therefore, recurring ways to describe valuations of many kinds, but are once again relational and constantly performed. In relation to data valuation, such registers might be connected to the knowledge emerging, though perceptions of what that means can vastly differ, ranging, for example, from trustworthiness to novelty or level of detail (Wagenknecht et al., 2024, p. 69). The comparison between data and tomatoes shows another feature of these registers. As value is relational, so are the categories used to describe it. Working with registers of valuing as an analytical tool, therefore, requires a two-step process, including the identification of shared ways of valuing and their analytical embedding into practices. The subsequent empirical chapters will follow this structure.

Coming back to the relational nature of valuation, I now want to explore what such relations might be; what features shape them? And how are they sustained over time? While it is established that valuations are relational, the extent of this can best be seen in cases where value shifts between contexts. Studying waste, Greeson et al. (2020) examine how value is dis- and reassembled. In their description, waste objects are seen critically in a number of ways, ranging from economic to cultural registers. However, when engaged with closely, objects described and valued as waste can be re-valued and put into new relations. This context shift comes with new valuations, removing the description and reintroducing new ways to see an object. This shift is enacted quite literally by practices as well, as something might be taken from a trash pile and recycled or reused, reintroducing it into a new function (Greeson et al., 2020, p. 159). In this case, the context of the trash pile itself is

influencing the valuation of the object, providing yet another way of looking at value. As part of a certain constellation, objects might be valued similarly, regardless of the person valuing them.

3.2.2 Valuation Constellations

The *valuation constellation* as an analytical tool introduces additional notions of structure into the description of valuation. Formulated by Waible et al. (2021), this approach aims to assemble consistent and recurring relations as stabilized circumstances for valuation and introduces additional vocabulary, especially for the description of practices. Highlighting the relational nature of value, the authors introduce *positions*[1], *rules* and *infrastructures* as collectively fixated. While not inherently stable, they can be stabilized and impact valuation across time and space. This means that valuation constellations need to be analyzed in terms of their effects on valuations. Certain positions, embedded in larger contexts, might emphasise valuations of certain kinds. For example, an editor at a scientific journal might evaluate a research contribution in alignment with their position (Waibel et al., 2021, p. 41). As part of their relational context, certain registers of valuing become relevant – like, quality of the research, fit to the journal and more – while others become secondary. This way of valuing is not just tied to this individual editor, but to the position itself, though related to the journal, as topical fit would not be the same for every publication. Extending this example, such a position is not only relevant when valuing others, but also when being valued oneself. As prospective readers of the journal make their judgments, the editor shifts from being the valuator to being the valuee. Once again, their position is shaping valuation, as it is most likely the editor's selection of research that influences the journal's valuation, rather than their work outside of this position. Of course, the two are interwoven as well. As editors might try to serve a certain community, their valuation shifts with the trends and interests they perceive. Similarly, in the context of biodata resources, users, curators, or maintainers might share perceptions and enactments of value with peers in similar positions. Nonetheless, it is also important to consider their valuation in relation to their contexts. While researchers in one field might find the data aggregated under a resource especially valuable for their work, users in a different field might not value it in the same way. What they might share is the perception of a resource in relation to its usefulness for research; yet, as their research differs, the valuation of data may yield different results (Pinel et al., 2020, p. 191).

Similarly, curators, making data movable, could share ways to evaluate and valorize data. Their selection processes are guided by evaluations of what might be useful elsewhere and what can be made movable (Pinel et al., 2020, p. 179). However, their selection is not just tied to their position and contextual entanglement. Valuation as a curator is tied to rules (Waibel et al., 2021, p. 42). In their position, curators select and interact with data in structured, stabilized ways. Remembering the accounts of Plantin (2018), curators are tasked with seeing data through a rather narrow lens. Certain

[1] Positions can further be differentiated from identities, which represent the internal perception of the position. This distinction is omitted here as it adds additional complexity and offers only limited analytical benefits to this thesis. See more details in Waibel, Désirée, Peetz, Thorsten, & Meier, Frank. (2021). Valuation Constellations. *Valuation Studies*, *8*(1), 33-66. https://doi.org/10.3384/VS.2001-5992.2021.8.1.33-66

features are identified and subsequently tied to the data as metadata. Data are evaluated and valorized with regard to these defined features.

> *Procedural rules enable and constrain valuations; they prescribe the conditions for engaging in valuation, define valuation criteria, formulate the proper steps for it to proceed, and specify how judgments are communicated.*
>
> *(Waibel et al., 2021, p. 42)*

Working in union with curation manuals, curators value, while embedded in their context, their stabilized role, and the stabilized rules[2] guiding them. These rules, whether formalized or not, are still tied to a certain context of community, meaning that, while they standardize valuation to a certain degree, they might not be compatible with judgments in other contexts, by other positions and under other rule-systems. Valuation, even when tied to rules, does not necessarily travel (Meier et al., 2017, p. 45). Like data, valuation needs to be made movable to be relevant beyond its context. Something that requires collective work. As rules are developed collaboratively, they enable the valuation across contexts and positions, providing a stabilized shared understanding. In large constellations, however, such discursive stabilization is not always possible, as entanglements change and positions are taken up by new members. Once the old rules are no longer collectively agreed upon, a new set of rules might emerge, putting into question previous efforts to produce consistency. Practices change, and with them the validations that underpin them.

To assure consistency, stabilization might therefore be delegated to valuation infrastructures (Waibel et al., 2021, p. 45). Emerging out of closed negotiations, these infrastructures configure future practices, stabilize positions, and thereby prescribe valuation. Standardized user interfaces, ontology-focused databases, or predefined visualizations travel from context to context, shaping the valuation of curators and users alike. At the same time, they define what both of these positions even mean by allowing them certain kinds of valuation. While the same person might be at times a curator and at times a user, the infrastructure only allows the enactment of one of their positions at a time. By being tied to a number of contexts and potentially linked to shared rules of valuation, these infrastructures allow for valuation to become mobile without explicit ongoing negotiation. Yet, valuation infrastructures do not just emerge. They are themselves created through extensive practices, underlined by values.

As developers and maintainers of infrastructures configure their design, they presuppose future practices and the valuations that come with them (Waibel et al., 2021, p. 46). In this process, they are enacting their own valuation of what those practices should look like, what positions could be taken by future users and what rules are best used in their valuing behaviour. The curation of data as a

---

[2] In their theoretical exploration of valuation constellations, Waible et al. identified three types of rules that can shape valuation independently or in union. As the valuation rules are only a peripheral element of this thesis, further exploration of different types of valuation rules is cut short at this point. The rules described in this thesis most closely resemble what they call “procedural rules” (2021, p. 42). For a more detailed account, see: Waibel, Désirée, Peetz, Thorsten, & Meier, Frank. (2021). Valuation Constellations. *Valuation Studies*, *8*(1), 33-66. https://doi.org/10.3384/VS.2001-5992.2021.8.1.33-66

valuing practice is then not just connected to the individual, their role as curators, their context and the rules agreed on. Curation is similarly predefined and shaped by the digital infrastructures in which it is embedded and which allow for certain practices while omitting others. Across contexts, the values embedded in the infrastructure might differ from those of the context. This can especially be the case if usage contexts are not well enough considered, or wrongly perceived in the development process. As usage practiced and usage envisioned diverge, the infrastructure might fail as a stabilizing device (Forseth et al., 2019). As such, those valuing are not without agency in the face of valuation infrastructure, able to push for changes in the infrastructure or to retract entirely. Due to the ongoing nature of maintenance required for digital infrastructures, maintainers' practices of development become entangled with their users' practices in an effort towards a more usable, *good* infrastructure, stabilizing more than their valuation.

Rooted in care for data, data work can be seen as a practice entangled with valuation. Valuations of *good* data, meaning data that is worth being cared for and data that is worth working towards, differ contextually, subjectively, and temporally as they are shaped by their entanglements. *Good* data and *good* data practices are perceived at all times in multiple ways, not just between individuals. To better grasp this multiplicity, registers of value can be identified as common themes alongside which valuation is structured. These are not specified ways of valuing data as "good", but instead describe recurring perspectives on data of any kind. Just as valuation is embedded, so are these registers tied to the contexts and the associated needs. In the first section of the empirical chapter, I will explore exactly such registers of valuing in relation to data, data work and the digital infrastructures that shape them. These will serve as an embedded theoretical lens for the subsequent exploration of practices, extending the concepts explored in this chapter.

Building on valuation constellations as a second analytical tool allows for a more nuanced look at the ways that valuation is structured across positions, in relation to rules, and when embedded into infrastructures. These positions enable a more structured look at commonalities between valuation practices, highlighting potential shared registers independent of context and individual. Tied to practices, this lens makes visible how data are seen differently, yet in repeating patterns across different types of data work. Bridging between positions and contexts, rules can be yet another way of describing how valuation is stabilized across time and space. Especially in relation to curation practices and shared ontologies, such an exploration of rules can further inform the long-term enactment of *good* data and, especially, *good* data work. Seeing this data work as embedded in infrastructures allows for the description of yet another level of stabilization. As (en)coded contexts that move between research fields and usage situations, biodata resources can be described in regard to their capacity to structure practices and valuation. They themselves are entangled in valuation practices, shaped by larger constellations and a need to be constantly maintained to enact good data and good data work.

In the second section of the empirical chapter, practices will be explored in relation to exactly these features. Their embeddedness in constellations of positions, rules, and infrastructures will be the

primary focus of my analysis. Here, the identified registries of valuing act as an embedded conceptual framework, providing emergent descriptions of what is and what is not considered in the valuing of data in biodata resources. However, this analysis, the description of values and practices, needs to be seen as contextual too (Doganova et al., 2014, p. 92). In the role of the researcher, looking from the outside, my own context and entanglement are at play. While my practices are aimed at the representation of the data work explored, they are also shaped by ideals of structure and narrative, to name just two. The identified register of valuing, their assignment to practices and placement into constellations, is as much a representation of my position as it is an account of the field I am entering. In the following chapter, I will therefore further expand on my approach, introduce the research questions guiding my approach and give context to my forthcoming analysis.

4. Exploring Valuation

This thesis is embedded in larger efforts to study the socio-technical entanglements of biocuration at the STS department at the University of Vienna. My engagement with this topic is tied to employment in the group *Technosciences, Materiality, & Digital Cultures*, engaging with the role of data and material practices across a number of sites. Previous work on biocuration (e.g., Davies & Holmer, 2024) has engaged with biocuration as hidden work and as a differently structured career than most academic positions. This work is prominently featured in my reading of past literature and has shaped the initial stages of the project. In this chapter, I outline my approach, acknowledging how the transition from a non-thesis project to the current format has shaped my research and possibly influenced my representation of the field. As such, I want to start with my own positionality before giving a more general account of my empirical work and the performed analysis.

Building on a media studies, sociology and political science background with a personal focus on digitization, I am entering the field of biological research data work as a social scientist. My previous projects focused on the inner workings of platform companies as socio-technical assemblages and the formatting power that these platforms have. This background shapes this thesis. Departing from the aforementioned work on biocuration as a process, I entered the project asking what these processes aim to produce and how biodata resources communicate their curation as “good” or “pristine” (as described by Plantin, 2018). Guided by this interest, two initial interviews served as an exploratory engagement with the field. The conversations focused on markers for quality in the doing and usage of data curation. Building on these interviews, the project transitioned into its current form, focused on the valuation of data and data work. While slightly different in their structure, these initial conversations still contributed to the project and remain, albeit to a lesser extent, a part of the analysed corpus. Their integration will be discussed in more detail below.

As a master’s project, this thesis sets out to explore valuation across a number of positions, along with practices in and around biodata resources. While stabilized constellations, like positions, rules and infrastructures, play an important role in this text so far, and beyond this point, their presence was largely the result of their emergence out of the material. Targeted at registers of valuation, and shifts,

negotiations and combinations of such, maintainers were initially identified as central figures in the building and changing of data work. Their management of more than just practices, including value constellations as a whole, has since become an equally important aspect of this thesis and can be seen as a result rather than a part of the project's expected scope. Beyond the practices explored, this project itself is designed to embrace relationality, making room from the start for such emergences, even on a structural scale.

The result is a methodical set-up and a set of research questions that explicitly highlight the messy nature of data work (following Law, 2004), while aiming to identify the values that are enacted through practices of any kind. Embedded in varied kinds of data work, valuation shapes how data are seen and worked on, evaluated and valorized. In this thesis, I, therefore, want to explore these valuations and the constellations that shape them. The chapters so far have established that looking at research data means looking at their enactment as such. Bound to their context of creation and the re-interpreted necessary for travel, data do not just present themselves for re-use. Data have to be made movable, all while keeping them *good* or *usable*. Through data work, this process of transport is enabled, as data and descriptions about their origins or potential uses are made translatable from one context to another. This work of overcoming distance is rooted in practices of many kinds, some performed in close contact with the data, others concerned with building the digital infrastructures housing them. Yet what travels are not just data; moving with them are enactments of their value, making the move worthwhile. Going forward, I hope to make them visible.

RQ: How are data valued by those working with and on biodata resources?

In exploring this question, my aim is not just to provide a summary of registers of valuing and practices, but to give a structured account of their interaction, as valuing is enacted in a number of distinct ways. As such, registers of valuing will first be identified and systematized before connecting them with data work of many forms. Here I am building on additional scholarship, highlighting the distinct work of maintenance, both in relation to software generally and in relation to biodata and open science resources specifically. This scholarship allows for a more nuanced understanding of practices across larger constellations, with positions, rules and infrastructures shaping what data work is and who it is performed by. Building on the identified registers of valuing allows me to distinguish how practices of different kinds are enacting valuations of different registers both within and in proximity to biodata resources. It is this work that is at the center of my analysis and will hopefully enable me to answer my additional research questions:

SQ1: How are data valued differently through practices in and around biodata resources?

To move beyond scholarship on caring data work and on accounts of maintenance processes, I want to further sharpen my lens once more. Once valuation emerges as stabilized and structured by constellations of many kinds, it becomes important to explore what is and what is not enacted in practice. As positions might clash with rules and are shaped by infrastructure, value is not just enacted

but is constantly being combined and negotiated. To examine data work in the context of digital infrastructures – with positions such as maintainers, creators, and users, and rules that shape more than just ontological description – means identifying where registers of valuing intersect and how their connections are made fruitful in practices and constellations to come.

SQ2: How are registers of value combined and negotiated by those moving through, interacting with and maintaining biodata resources?

My approach to answering these questions promises to provide a deep understanding of the dynamics at play, as I interview people who are active across many positions in comparatively small-scale settings. As such, this thesis ends up doing more than I initially set out to do, identifying registers of valuing, relevant practices, as well as their underlying positions, rules and infrastructures. As these emerge, the planned ethnography of data work and value moves closer to what Star (1999) calls an "Ethnography of infrastructure." In exploring practices, this thesis follows her call to explore mundane practices, breakdowns, and negotiations to make visible what is otherwise hidden. Yet, in this case, infrastructure is not just perceived from the outside. As maintainers talk about changing structures and the stabilization of valuation, their work of infrastructuring provides a perspective that usually remains hidden. While predating Star's methodical exploration, Ball's book chapter "Researching Inside the State" (1994) provides a fitting addition to describe the interviews conducted. As such, maintainers' accounts are not just interpreted in regard to their practices but also under consideration of their stabilizing power (Ball, 1994, p. 114). This is especially relevant when describing the emerging valuation constellations.

The thesis mainly describes resources with narrowly focused collections that are serving specialized fields and expert communities, which brings a number of affordances for the analysis. As small resources are often overseen by individuals or small groups, interviewed maintainers were anticipated to recreate a representative image of their resource and its development out of personal accounts. As key decision makers, they are deeply involved in the navigation and negotiation of divergent value representations. Beyond that, as teams are small, positions are less likely to be strictly divided between people, enabling interviewed maintainers to give insight into the different positions that they need to enact within their person. Yet, maintainers are not the sole focus of the empirical strategy. To take into consideration supplementary accounts, the additional groups considered for the sample similarly bridge contexts. Combining the perspectives of outsiders and contributors, community-curators were selected to provide a counterpoint to the internal perspective. Their work is closely involved with the research data themselves, that is, in their positions as data producers, as data curators and as data users. Lastly, members of larger consortia, collectives of researchers and stakeholders, append the sample. This group, while also embedded in data work and maintenance activities, brings an outside perspective due to its extensive connections and bird's-eye view of the field.

4.1 Finding Valuators

The first two interviews, initially serving a primarily exploratory purpose, were selected from a pool of people already known to the research group in which this project began. Still, both fit the sampling strategy discussed above. The two interviewees work on the maintenance and design of data resources and data work tools, shaping the curation performed within their more or less institutionalized operations. Based on their representation of the field and the sampling strategy outlined above, additional research was conducted into the relevant biodata resources, identifying several suitable small to medium-sized resources with community-involved curatorial practices. Based on this list, contact addresses of resource managers were identified, leading to three additional interviews with maintainers. While a similarly direct approach was initially planned for community-curators, contacting curators mentioned throughout collections yielded no results. As such, the two selected and interviewed user-curators were identified through recommendations by previous interviewees. Lastly, three relevant consortia were identified based on the previous interviews. These are groups of curators, maintainers, and other stakeholders who oversee certain aspects of data work and management across individual biodata resources. Their work is driven by the motivation to bridge potential gaps and establish standards. Outreach to board members of these consortia led to another two interviews, completing the sample of nine people. As all people interviewed serve multiple positions and contribute in various ways to data work locally and beyond, this thesis is building on a very engaged community.

Apart from their positions and responsibilities, interviewees are all far into their careers, with at least finished PhDs and often years of experience in their current roles or the contexts they operated in. Among the sample are five women and four men, spread across Europe in four instances and the USA in five instances. This distribution appears to be in line with larger distributions within the field; however, the geographical focus of biodata resources on Europe and the USA, especially, was problematized among the interviewees and is similarly a weak point of the assembled sample. The lack of diversification in this regard is in part due to the selection process based on available lists of resources, which are themselves positioned within these geographies. In cases where databases outside of Europe or the USA were identified, a lack of contact information or language barriers hindered the outreach and planning of interviews.

Table 1: Interview Sample

Person	**Roles & Tasks**
Louis (M, DE)	Consortia work and maintenance of a community-managed & -curated resource.
Bejan (M, UK)	Building tools for expert-curation at a large-scale resource.
Hannah (F, UK)	Maintains and moderates a resource using researcher-curation.
Jan (M, US)	Maintains and moderates a resource using researcher- & user-curation.
Valentina (F, US)	Maintains a resource using user-curation and user-moderation.
Liv (F, US)	Researcher, user and community-curator.
Luka (F, US)	Researcher, user and community-curator.
Nico (M, UK)	Consortia work and maintenance of a resource with expert curation.
Sophie (F, US)	Consortia work and maintenance of a resource with expert curation.

The summary in Table 1 provides a rough overview of the people interviewed, though it omits identifiable details and changes the names. As stated before, most of the resources that interviewees are involved in are community-curated, meaning that large parts of the curatorial data work are taken over by researchers and users of the resource. Despite this commonality, their approaches to curation, available resources, and priorities all vary. They differ in who contributes and in how community curation is managed. It is therefore important to note a few of the differences described in the "Roles & Tasks" column of the overview. One resource mentioned is strictly curated by researchers submitting their own research, while two others solely rely on users curating whatever they deem useful or relevant. A fourth resource aggregates data curated by both researchers and users. Similarly, the moderation of curation is handled differently, with some resources reserving moderation strictly for maintainers and others offering the possibility of user-moderation after a status promotion. In three cases, resources rely on expert curation; however, these curation efforts were only in part relevant during the interviews. Other activities of the interviewees, such as involvement in consortia work, were more relevant.

4.2 Talking Valuation

All nine semi-structured interviews were designed to both focus on relevant topics while allowing for the emergence of new relevant elements. These interviews were conducted based on guidelines tailored to the respective position (Galletta, 2013, p. 24). To ensure both consistency and a fit, a total of four interview guidelines were developed: one for the exploratory interviews and three quite similar ones for maintainers, users, and those overseeing the field as members of consortia. As the initial two interviews differ slightly in focus, their interview guideline only shares larger themes with the subsequent ones. The first two interviewees were asked about their perception of the field at large, possible descriptors for biodata resources and potential markers for determining curation quality. Further questions concerned their past and current work, as well as relevant stakeholders involved in changes to the data work they conduct or facilitate. For the remaining seven interviews, guidelines remained largely consistent, with minor differentiations for the positions held.

In general, interviews were structured in three phases in line with Galleta's (2013) approach to semi-structured interviews: The first concerning the immediate work of the interviewee, their path to their current position and the important factors influencing their day-to-day work. In the case of community-curators, this involved relating the resource to their research activities, while maintainers and members of consortia were asked to specify their managerial responsibilities. The following second section focused on curation as a practice more generally, initiated by asking about the way curation is or should be handled at the respective organisation or at large. Once again, historical changes and the underlying decisions were an important aspect of this section, especially in the case of maintainers whose decisions shape the data work of others. In the case of the user, this section focused on changes in their usage over time. For interviews about the larger context, this section aimed to bring out opinions on different types of curation, their advances and drawbacks. The third and final phase of the interviews provided the opportunity to discuss the topic of the thesis directly.

Opening this section, interviewees were to be informed about the valuation studies approach of the research (very broadly) and asked to speak to notions of value in their own work, resource or the field in general. Once again, for maintainers, this focused on managerial tasks, especially in relation to funding and impact. Similarly, with the group actively engaged in consortia, funding and impact were discussed on a larger scale. Slightly shifting the focus, users were asked about the value that resources provide for their work, while also leaving room for considerations of impact and funding.

All interviews were performed over Zoom and recorded using both the internal recording function and a backup software running in parallel. As part of the correspondence beforehand, every participant was informed about the recording process and further pseudonymized use. To ensure that consent was given for further use and publication, all interviews started with a standardized two-step process. First, interviewees were asked for consent before starting the recording, both within Zoom and in the backup software. Following this step, all interviews began with a standardized text outlining the terms for recording, further uses, and the interviewees' rights to redact part of the interview or withdraw it completely. Once again, interviewees could consent before the interview began. The first two interviews, initially not meant for publication, followed a slightly different approach, where consent was given at a later stage via email after their inclusion into the analysis became relevant. These two interviews were conducted in September 2024, with the remaining seven interviews taking place in February to April 2025. Interviews lasted between one and one and a half hours, depending on the conversation and the availability of the interviewee. The conversations were conducted in English, and every interview went smoothly and as arranged. As interviewees talked about their experiences, relevant topics that extended the scope of the planned questions were identified and, if possible, folded into the following questions (Galletta, 2013, p. 24). The result was rather a natural conversation, guided by both the planned structure and the specific experiences and perspectives of the interviewees. This approach was successful, revealing additional information about stabilization and structures, alongside the relevant ways of valuing and the practices involved. After completing the prepared parts of the interviews, all participants were given the opportunity to provide further comments, in case any topics relevant to them had been left out. In some interviews, this sparked further insightful discussions, while one interviewee took up the offer to send additional remarks via email. Both the closing comments and the email are part of the analyzed corpus. The email is marked as such when cited.

Following the interviews, all audio recordings were stored locally and removed from the Zoom platforms. Using two AI transcription software, Wrealley Transcribe and Descript, each with strong privacy policies, all interviews were transcribed in two phases. A first automatic transcription provided a general structure and timestamps to the text, before additional manual curation removed potential errors, filtered out filler words and anonymized interviewees. The resulting transcription was once again stored locally, and all audio files were removed from the transcription services.

4.3 Analysing Valuation

Analysis for this project followed an abductive approach (Vila-Henninger et al., 2024). Guided by the theoretical foundation of registers of valuing, yet open to emerging themes – or in this case structures. Analysis was conducted using the software MaxQDA, following a multi-step process. While initial notes were collected during the editing of transcripts, these only partly influenced the subsequent coding. Starting with an initial round focused on the different ways that data, data work and biodata resources were related to value, codes were identified based on the text. After coding two interviews, the resulting code book included codes, clustered into categories like "conflicts," "benefits," "actors", and "descriptions of data." While initially helpful, the application of this code system proved challenging, leading to a restart. This time, the coding was aligned more closely with the theoretical considerations laid out above. Coding both registers of valuing and practices, a second coding schema emerged. While mostly focused on these two categories, some additional codes were still captured, leading to a temporary third category of conflicts. As codes of this third category proved scarce, it was discarded despite opening potentially valuable analytical possibilities. While the content of these coded segments is still present in the final analysis in the form of quotes, a systematization was not possible, probably due to a lack of questions directed towards this aspect specifically. After coding the entire corpus twice with the new coding scheme, additional categories emerged within the clusters of registers and practices. These sub-categories represent what is described above as emergent structures. Throughout the interviews, valuing could be aligned along a number of registers. Interviewees discussed data in relation to their knowledge value, usefulness, and usability. This formed the most important categories of codes. While each combines more nuanced distinctions, the three clusters form the primary registers of valuing. Referencing the epistemic nature of data, most interview sections included codes under these categories. An additional four categories were identified throughout the texts, referencing more than epistemic questions. Next to the primary registers, these secondary registers of valuing will be explored in more detail in the first empirical section of this thesis.

While the analysis was initially geared towards practices of data work more generally, interviewees throughout the corpus made distinctions between three types of practices. The final coding structure, therefore, recognises practices that are performed within the data pipeline, including selection, curation and communication of data as well as practices by people outside the resource and those performed by maintainers. Those interacting with the pipeline from the outside are now represented in codes describing the publication of research, the contribution of time and the usage of aggregates. Lastly, maintenance practices are represented as maintenance itself, coordination with larger bodies and peers, as well as context creation in the form of funding or institutionalization, performed by third parties. Beyond each of these practices are more specific ways of doing each of them. They will be discussed in the second empirical chapter of this thesis.

Before moving on to these two chapters, it is important to note the necessity for this two-fold approach. The identified registers of valuing serve as more than just empirical results. While

insightful in their own right, they are used in this case as an extension of the conceptual framework discussed above. As contextual categories that make describable the highly contextual valuing, they allow for an easier description of valuing in the analysis of practices. Structured under registers, the coexistence and conflict of ways of valuing will be made visible in the second empirical chapter. This approach, while lengthy, provides the needed structure to an already complex set of registers of valuing and a similarly extensive exploration of enacting practices.

Following this approach, I hope to sensibly condense more than nine and a half hours of interviews in a way that contributes to the understanding of data work as a distinct situational practice that is shaped by valuing and its constellations. Answering my research questions, the theoretical concepts selected will help make my exploration of maintenance in the context of biodata resources fruitful. And in closing a long research journey, the methodological approach laid out above will allow me to contribute to the scholarship on data work, maintenance and biodata resources going forward.

5. Valuing Biocruation

Starting my empirical analysis, I will set the practices aside – just for a moment. The aim of this first analytical chapter is not to understand what users, maintainers and all the others involved in biodata resources do. Instead, I want to build an analytical toolbox, extending the theoretical frameworks that I already laid out. By collecting the different ways to make sense of data, of curation and of the biodata resources in general, the aim is to understand what is valued by those involved. While empirical in nature, this will produce a second order of theoretical concepts that are fitted closely to the site of my research. These resulting registers of valuing will later help look closely at practices in and around the data pipelines. As discussed above, registers are not concrete requirements or sentiments voiced in relation to data, curation or the aggregated outcomes. Instead, they are categories of thinking that emerged in the interviewees' accounts of their work – sometimes explicitly, sometimes less so. Among them are valuations of knowledge, its usefulness for certain purposes and the usability and affordances of the resources as technical infrastructures. These I will explore in detail in the next section. Beyond these primary ways of valuing, a secondary layer emerged. These registers are not directly linked to the data and their use. Instead, interviewees mentioned valuing non-epistemic uses, personal and community benefits, and brought up larger questions of openness and control. These secondary registers will also be explored in more detail in a subsequent section. The result is a tailored set of theoretical concepts that extend the more general considerations in Chapter 3. Embedded in place, yet simultaneously rooted in a valuation studies approach, these registers of valuing help better describe how maintainers, users and other stakeholders make sense of data and data work.

Only after establishing these registers of valuing, I will focus on the practices involved in 1) working within, 2) interacting with, and 3) the infrastructuring of biodata resources. Each category of practice will be explored in detail in the sixth chapter, while making connections to the registers of valuing identified. Here, conflicts that arise and negotiations that need to be made will become important. The

interplay of the different positions involved becomes apparent. This two-step approach allows me to make explicit what is relevant in the practices identified, while highlighting the position of certain groups of stakeholders, their actions and their ways of making sense of data aggregation. As ways of valuing flow through the practices in a number of directions, they need to be integrated and negotiated again and again. This approach will itself become crucial to making them visible.

5.1 Primary Registers of Valuing

> *It makes no sense for individuals to do this when it should be a collective resource. So, I think the value then, the value of the combined data is infinitely more than the value of just subsets of the data anyway. And the fact that we can combine it in any way, and that we can immediately get an overview of how much we know about the organism, really, we can kind of break it down into sections and go, these genes do this, these genes do this, and […] we use a process slim for this. So we can kind of divide the biology into different processes in the cell, and then we can go, right, which genes aren't annotated to any process, and that's valuable. That's kind of emergent knowledge, I guess.*
>
> *(Hannah, Pos. 68)*

Segmenting registers of valuing into primary and secondary is mostly for analytical purposes. But the selection of each category is not arbitrary. The primary registers are those that always arise in close relation to the data, the curation and the biodata resources themselves. Often, they are the explicit reasons for building, maintaining, contributing to or using a resource. In Hannah's account, they all find a place. In this section, I highlight all three of these primary registers of valuing and define important factors influencing them. To get an overview, I want to look closer at the quote above.

The first and most central register of valuing is the "emergent knowledge", enabled by "combined data" and providing an "overview" of organisms and biology at large. More bluntly: "the value of the combined data is infinitely more than the value of just subsets of the data anyway." Knowledge and aggregated data are not surprisingly important aspects, but not everyone agrees on what constitutes "good" knowledge and data. Important factors are whether the data are complete and correct, the way they are structured, and their trustworthiness. The second register is the usefulness for academic research (and downstream causes). Described here as the "breaking down" of organisms and the ability to identify "which genes aren't annotated," this connects the knowledge with the needs of academic research, and in other accounts, downstream societal impact. This category is often a measure of success for biodata resources, though not necessarily one that is easy to measure. Lastly, time saving and the affordances of digital resources stand between the knowledge and its use. "It makes no sense for individuals to do this," describes perfectly the logic behind a shared data resource in an academic system with limited resources. But these resources allow for more than timesaving, offering, for example, the option to apply a "process slim," a feature for categorizing genes into biological processes. Usability is, therefore, also a matter of technical affordances.

In the next three sections, I will explore the ways that these registers are brought up and what they entail. The result will be an embedded vocabulary allowing me to further analyze how value is made sense of in relation to data, and how this valuing is enacted through practices.

5.1.1 Knowledge Value

> *We curate functions, processes, components, phenotypes, modifications, expression levels, and a few other things that are less important, and we try and do that to completion. And we don't know that every single piece of information is going to be useful, but we don't want to have to go back to papers later, necessarily and capture things that we missed. So we capture everything, and we hope that somebody somewhere is going to just find some piece of information useful.*
>
> *(Hannah, Pos. 26)*

Between resources, there can be quite a difference in what is meant when talking about knowledge. In the case of knowledge bases, like the model organism database maintained by Hannah, the aim is to gather everything known about a certain organism. But the question remains, what does "we capture everything" mean then? In her case, research papers are entered into the system with as many details as possible. In another case, a clinical knowledge base focusing on cancer research, the selection is a little bit different:

> *So the idea is really to help researchers and clinical teams to identify relevant literature for […] a clinically relevant association between a variant and a cancer type.*
>
> *(Valentina, Pos. 50)*

Here, the knowledge is valued in relation to relevance. Papers with relevant information to certain fields are selected, knowledge is extracted and then integrated into the larger collection. With each addition, the aim is to fill gaps, to provide more detail and to correct previous errors. What exactly is meant by relevant will become clearer in the next section on usability of biodata resources. What is clear already, however, is that more data is not necessarily better.

> *How do you both make a comprehensive database, something that has all the information, while also making it consumable? Something that, when a person goes there, they don't feel overwhelmed, and they can actually find the relevant information. It may be that there are 20,000 fluorescent proteins, but only a hundred of them are really ever used frequently. And you need to somehow not bury the meaningful information in the archival information, so to speak.*
>
> *(Jan, Pos. 99-100)*

The question of complete knowledge is therefore one of fit to the requirements of the field and its stakeholders. As such, the ideal level of detail in the curation is a contested factor, not always easy to determine. This becomes even more challenging as resources try to make their data relevant both to

current and future research. Navigating the needs of researchers and predicting how those requirements might change, the selection of data to include is a constant consideration. But completeness of knowledge is not the only aspect considered. To serve their purpose, biodata resources need to be filled not only with enough up-to-date data.

Data, and especially their curation, need to be correct. As more and more data are aggregated, small errors might accumulate, potentially leading to inconsistencies. Making sure that everything is correct can become relevant in a number of ways. To make sure that the data populating the resource are valid, biodata resources tie each annotation to its original source. This so-called *provenance* information not only provides further insight into the context of data creation, but it also ensures that information entered into the database originates from peer-reviewed research. The publishing status of research alone is not an assurance of correctness. Still, it is often an indicator of what can and cannot be included. Biodata resources, in that sense, act as a reflection of the research field. However, the originating research is not the only element that can be faulty.

> *Basically, the science can be wrong, the curator can interpret science incorrectly and curate it incorrectly, or the ontologies that we use to describe the data can be wrong; they may have things, terms misdefined or placed in the wrong place in an ontology. So if you use them, you get the incorrect information. So that's at least three ways. And then we have inferences from knowledge that's transferred from other species and that can be wrong.*
>
> *(Hannah, Pos. 20)*

As resources combine research from different fields and subfields, these inferences can pose a challenge to the correctness of the data. This happens both in cases where data is taken from different strains of the same species and especially in cases where knowledge is inferred between different organisms.

> *There are so many strains, like all these different strains... […] so that also begs the question of, when they're doing the annotations, […] is that reflecting what you see on one strain versus a different strain?*
>
> *(Luka, Pos. 92-93)*

> *There's a lot of inference, and you know, one of the worries about all of this stuff in general is that you have to be very careful to mark up what is an inference and what is a primary piece of knowledge where you have direct evidence. And actually, not all of our systems do that.*
>
> *(Nico, Pos. 131)*

While these concerns are less about the correctness of the knowledge, they are about the correct understanding derived from it. These different perceptions of what correctness means, what it is directed at and what its influences are highlight the multiple ways in which a certain register of

valuing can be applied in practice. This becomes increasingly complex as different nuances come together.

In the last two quotes, research provenance is not just a question of linking to the original research; instead, the interviewees highlight the importance of annotating in the right way. While their concerns are about correct interpretations, their valuing of knowledge is also about its structure. That is the way that a certain resource described strains, or inferences. How data are structured influences interpretation, and, as will become more apparent later, downstream use. Structuring data *well* means describing it in a way that makes it understandable beyond its original context. In relation to the previous quotes on inference, this potentially means that by harmonizing data under a structure…

> *you decrease the chance that somebody, in trying to pull the data together, doesn't really understand what the differences are and misinterprets it and brings it together in a way that leads to incorrect conclusions.*
>
> *(Sophie, Pos. 104)*

The structure used by a resource is not just about making sure data is interpreted correctly. As biodata resources aim to aggregate data from various sources in an attempt to make them less messy and less scattered, the structure of the data is often related to future use.

> *There is a format, like a data standard, called the PSI-MI TAB format, and that exists. It's very well described. It has a very simple structure, yet it's very detailed, and it allows you to put as much information as you're able to get. So if you're looking at a resource that says we're doing protein-protein interactions, and well, if they give you a two-column CSV file with some random names inside it, that's not so great. You can tell that the quality [of the resource] is very low because they haven't thought about "what are the standards we can use and how do we make this data really useful for someone."*
>
> *(Louis, Pos. 20)*

Providing useful data, though, is not only a function of having or not having a structure applied to the data aggregated. The way data are structured and the selection of standards they are subjected to are similarly important. As representations of biology into which data are fitted, ontologies both describe and are the result of the research collected. In mapping out the biology, they represent the current state of knowledge while constantly subject to questions of breadth, detail, and timeliness. Talking about the value of structured knowledge includes talking about how well that structure serves the needs of users and researchers. With changing needs, ontologies are, therefore, in constant change and subject to constant evaluation.

> *It doesn't necessarily mean the original annotation was wrong; it just means that now it's cleaner. And the redundancy's gone. And sometimes it is fixing things. […] It's not terribly biologically incorrect, but it's ontologically incorrect, and it's not great for logical inferences.*
>
> *(Hannah, Pos. 57)*

To allow these inferences – comparisons of biological features between different organisms – ontologies need to be consistent across fields, describing the same features in the same ways. This integration, however, is not only valuable on the logical level; it also allows technical integration between different resources that might, for example, describe similar organisms. This integration is itself a valued feature of the knowledge produced across biodata resources, providing research with more data. However, it also comes with challenges.

As data are collected from across fields and subfields, curated and annotated through resources, before being published as an aggregate on these websites, they are further and further removed from their origin. This means that the epistemic distance they need to travel increases. While provenance data provides a path for users to trace knowledge back, there is a need for trust in the entire process. There is a need for trust in the process of making data movable. In the case of community curators, for example, this trust is derived from their intimate knowledge of the internal processes. Through their own involvement in the resource and especially in the curation process, they can trace back how knowledge comes to appear on the resource. As such, they use the resource with more trust.

> *I value the community curation very highly because even if someone reads your paper once, they might not remember or look at supplemental figure five C or whatever, you know? Whereas, like, it's fresh in your mind, and I just feel like there's the ownership of your data; for me, it's very motivating to get it represented accurately in the database, whereas if it's someone else's work, that motivation may not be there.*
>
> *(Liv, Pos. 111)*

Such intimate knowledge is, of course, accessible only to a few involved members of the resource community. Yet, for others, trust is just as important in the interaction with biodata resources. Here as well, information about the aggregation and curation procedures acts as an important indicator, shaping not just trust in the resource but also its subsequent use.

> *I would value a resource and I would trust and have confidence in a resource that is transparent about who are the people doing the curation, more than one that doesn't say who did the curation. There's a lot of implicit things to unpack there.*
>
> *(Louis, Pos. 26)*

This unpacking is done in various ways according to the positions and contexts of those using and working on the resource. Valuing through the register of knowledge means considering completeness, correctness, structure and trust, but how each factor is interpreted and put into action looks different

for most. As the most central register, this understanding of knowledge in the context of biodata resources will become central when analyzing practices of many kinds. The value of knowledge is considered by everyone involved in their own way. However, knowledge is not considered alone. In the next section, a closely related register will be explored. Here, knowledge will be put into relation with its context of use.

5.1.2 Usefulness Value

> *The whole field is just people who really want to organize information to make it useful again.*
>
> *(Louis, Pos. 4)*

During the interviews, usefulness comes up a lot. All the work that is done to bring together knowledge is done for a purpose. But what does useful really mean? The better question seems to be: useful to whom and in what context?

> *Many of these [databases] are very, very specialist. Many of those are in, sort of, a long tail where they're probably very important. You know, some of the really specialist ones are probably very important for a comparatively small number of people. But we know at the other end of the scale, there's a comparatively small number of databases that are important for a comparatively very large number of people.*
>
> *(Nico, Pos. 53)*

The biodata resources at the focus of this thesis are most likely in what Nico calls the long end. Very specialist knowledge bases that serve comparatively small research communities. They are embedded in research fields and sub-fields, where they aim to aggregate the knowledge most relevant to their users. For people around the resources, they serve concrete purposes, enabling certain tasks of the day-to-day research activities. Being a helpful tool is a part of their value.

> *Throughout a project, we're gonna be continuing to use [the resource], but it really plays a critical role in the beginning stages of a [research] project, the planning stages.*
>
> *(Liv, Pos. 147)*

The academic value is, on the one hand, in the information provided. Biodata resources enable an up-to-date understanding of the current research landscape – of what is known and what is yet to be researched. They are valued for providing the information that is needed for current research. On the other hand, here, too, the way that data and knowledge are structured becomes important through standardization and integration with digital tools. Aggregated standardized data has been an integral part of enabling new methodologies, especially in bioinformatics. As information is standardized and made computer-readable, new possibilities arise. These large datasets of structured knowledge are seen as especially valuable in regard to a variety of computational methods, especially AI training and

use. One such case is the development of AlphaFold, an AI tool that visualizes protein structures based on machine learning algorithms.

> *And so then what happened with AlphaFold is you got an opportunity to use all of that really valuable, well-structured data from the experimental science. Use that to train an AI system to be able to predict new protein structures based on the patterns that are observed based on the input sequences.*
>
> *(Nico, Pos. 30-31)*

AlphaFold is an example of academic usefulness in two ways. For one, it is the direct result of data aggregation through biodata resources, though not just the kinds described here. The use of aggregate data for its development shows the value of structured data for new kinds of research. On the other hand, AlphaFold itself enables new kinds of uses for researchers. Comparing sequences and visualizing data in such new ways positively impacts research more generally, as access to resources can be useful in new ways. Beyond these practical matters, using data for research purposes is not just connected to academic needs. Throughout the interviews, the usefulness was also a matter of downstream societal impact.

> *We think of rare disease and human health in general, we think of biotechnology and the economy. All of these areas are now very directly impacted by the science that comes from using all these data.*
>
> *(Nico, Pos. 142)*

In describing these societal and ecological uses, interviewees rely again on terminology that highlights the features of knowledge and its usefulness. Talking about clinical impact and the time saved by having data aggregated and structured, Sophie concludes:

> *You're making it easier for the clinician who has a patient who has this set of symptoms to be able to say: Oh, okay, I've done that, I've done my genome sequencing, and now, I can put this into a tool and get suggestions for where I should go next.*
>
> *(Sophie, Pos. 52)*

The knowledge provided in biodata resources is valued, but so is its usefulness to academic and downstream uses. This second register enables a description of valuation regarding the impact that data have in their context of use. Between the two, however, sits another important context. Already mentioned throughout the previous quotes is the pipeline itself, the path that data take and that makes knowledge accessible to users. Digital tools and analytic website features are not only a new way to interact with data, pointing to new research opportunities or even enabling completely new methodologies. The social and technical infrastructures underpinning them are described by the interviewees as valuable in their own way.

5.1.3 Usability Value

In the introduction of this chapter, aggregated data and overviews of current research development are described by Hannah as "emergent knowledge", something that has a larger epistemic value than its parts. This value is only partially in the aggregate and its usefulness, but also in the ability to interact with it easily and conveniently. Valuing these resources and their data through the register of *usability* means talking about the affordances they provide. Their features, the ability to sort and search, for example, are not inherently connected to the aggregated data; in fact, most resources discussed in the interviews emerged out of shared Excel sheets. As described by one maintainer, these features need to be considered away from the data:

> *So, for some reason, the word value made me think of like added value, right? There's the value of the data itself, and then there's added value. Like, in addition to storing the data, making a little web interface. Like the spectra viewer I showed you.*[3]
>
> *(Jan, Pos. 178)*

Providing data viewers makes data accessible in different ways than their original form. Instead of research papers, biodata resources provide interactive interfaces to explore or directly find data. In addition to user interfaces, computer-readable interfaces, also called APIs, provide yet another way to interact with data. Such an API is aimed at users extracting data in its initial computer-readable form, allowing its use in bulk. Both visual tools and APIs are useful only to specific groups doing corresponding research. What is valued as *usable*, therefore, depends again on the user's context and needs.

> *For researchers who are trying to use data for [machine learning] or something like that, I am glad that they can save themselves a lot of time and just grab it [through the API]. But that is definitely a different use case than the one I originally started with, which is just having microscopist have a website to go to, to find their next protein.*
>
> *(Jan, Pos. 131-132)*

Beyond the affordances to individual users, usability is also related to the ability to collaborate. Through curation tools that are part of the interface, the data work of many people across localities is enabled. Similarly, integration between resources depends on compatible and integrated technical infrastructures. Digital affordances of the resources make connected research data a reality. As digital infrastructures, they allow movement across time and space. However, valuing usability is more than a question of added value for users and curators. For small resources, building on top of the data of larger ones, these affordances are a crucial factor in successfully providing a service.

---

[3] The Spectra viewer mentioned by Jan is a visualisation tool that renders a graph of the light spectrum a fluorescent protein can emit.

> *If you took these core databases away, then many databases that are more specialist, that depend on them for their data input, or for the tools, or for some part of their process, they would stop being viable.*
>
> *(Nico, Pos. 56)*

Of course, these resources rely not only on the technical affordances of the larger resources, but they also require the provided data of these general repositories to be valuable in the ways described above. Only then can they provide the academic usefulness to their respective community. The usability of data and data resources is the third primary register of valuing, completing the first set of embedded theoretical tools for the analysis later on. These three primary registers are intertwined in complex ways. All three need to be catered to, it seems, in order to make a biodata resource valuable. But their interaction is not always without challenges, as each comes with additional features, all of which are highly dependent on contexts, positions and practices. This will become apparent as soon as these registers are integrated into my analysis of the practices. Before doing so, I will explore a number of secondary ways of valuing in the next section. Different from the primary register, these are not as closely tied to considerations of data themselves. Still, they are ways that interviewees describe data work and biodata resources as valuable.

5.2 Secondary Registers of Valuing

> *I get a lot of enjoyment out of [the resource], but I also really should credit it with extensive support of my career.*
>
> *(Valentina, Pos. 196, in a post-interview email)*

Talking about resources beyond the data, the tone shifts. From academic usability, interviewees move to talking about personal benefits in and outside of academia, about community and larger ideals. These secondary registers of valuing are important to them, driving their engagement and guiding their decisions. For the purpose of this thesis, they allow a more nuanced perspective of resources, making visible a more personal perspective on data work. The secondary registers are not focused on data or curation, but more on the people – either the interviewees themselves, their careers and connections, or on others around them, their collaborators and funders. These sections of the interviews are about ideals as well, arguing for openness in relation to a necessary level of control. In total, four registers through which interviewees value the resources and the work around them emerged. The first register of valuing that will be discussed in this section still relates to the academic system, as interviewees talk about benefits that are not epistemic in nature. This means looking at biodata resources through the lens of citations. Beyond that, personal benefits of different kinds will be explored as a second register. They range from descriptions of learning to the enjoyment felt by Valentina in the quote that opens this section. Thirdly, community and collaboration are valuable in themselves and for the advancement of shared interests. Finally, the section is closed by questions of openness and control. A tension that seems to run along a lot of the decisions made by maintainers and users alike.

5.2.1 Non-epistemic Academic Value

Some don't know, some are sure, and even others find it unlikely. When talking about the impact that resources might have on the citation of papers that are featured in them, the verdicts seem inconclusive.

> *So it's kind of really hard to say that curation increases the visibility. We have a few ad hoc examples where we get a paper into the database, and then we let the community know via our mailing list, and then we can see a big spike in accesses.*
>
> *(Hannah, Pos. 31)*

Whether actually effective or not, the (prospective) impact on citations seems to be an important factor for users and maintainers alike. Without preempting the following chapter, this question of citations often comes up in relation to time commitments and community curation. But citations are not the only non-epistemic benefits that might make a resource more valuable to those contributing to it. Being connected to a resource that is important to a larger audience comes with credit of a different kind as well. Talking about her position as a maintainer of a resource, Valentina writes in an email following the interview:

> *It has also provided research outputs in the form of high-profile publications that I have been able to be first, co-first, or co-author on, with visibility to collaborators. I have also spoken about [the resource] at many conferences and have been an invited speaker due to my experiences […] on three continents – it has hugely impacted my international reputation*
>
> *(Valentina, Pos. 196, Email)*

Going beyond the individual perspective, resources themselves are talked about in terms of their academic success. This includes the impact on the research community, at times beyond the usefulness and usability mentioned above. As valuable knowledge and academic usefulness are not easily quantifiable, resources need to find ways to communicate their worth. Especially in relation to funding, the rather abstract notion of impact plays an important role in signaling the value of a resource.

> *I think the impact that we do have is huge, but we don't have a great way to measure it. […] The problem is, particularly for knowledge bases, a lot of people use us, but they don't cite us. We're working on that at the moment by just reminding people, "You know, you should cite us here."*
>
> *(Hannah, Pos. 70)*

While not directly related to the data and their curation, these non-epistemic ways of looking at resources seem to play an important role. As such, they provide another valuable lens for the analysis of practices. The importance of this register will become especially apparent when looking at the coordination between maintainers and funders later on, though not only there. Both on the personal

level and for resources as their own entities, valuations are linked with measures of impact in different forms.

5.2.2 Personal Value

Away from academic measurements, biodata resources are still valued by the people involved. Here, again, careers and career opportunities are important factors, but not the only ones. As described by Davis & Holmer (2024), working as a curator, for example, can provide an environment quite different from other positions in academia. While time-consuming and often underrecognized on a larger scale, working as a biocurator comes with its benefits.

> *The reality is that curation is, kind of, a very great field where you can work remotely. You can work part-time and still be impactful, scaling with however much time and motivation you're willing to put into it.*
>
> *(Louis, Pos. 70)*

But time spent working on curation is not just described through the lens of working conditions, especially as many of the resources featured in this thesis mostly rely on community contributions. Interviewees talk about learning ontology terms and gaining knowledge extraction skills that come as part of the curation process. While some might question the relevance of the knowledge gained, others perceive the skills required as part of the professional growth process:

> *I think [my principal investigator's] and my idea align really well. We both think it's a really important step, and I think when I was a graduate student, she even viewed it as a really important training step for graduate students to learn how to do this. And so I definitely learned to value this from her.*
>
> *(Liv, Pos. 122)*

Beyond these benefits for personal development, maintainers and curators alike describe their experiences with words of pride and accomplishment. Curation itself, while described as tedious and time-consuming, provides closure to community curators who are putting their finished project into the database.

> *For me, a lot of my sense of accomplishment comes from that paper being curated on [the resource].*
>
> *(Liv, Pos. 143)*

Similarly, maintainers like Valentina describe their own feelings of pride and accomplishment. For them as well, their biodata resources provide an emotional value. This register is evoked especially when talking about the amount and quality of data aggregated, and the usefulness for users and research communities.

> *Hearing somebody else talk about it: […] "Oh, we use [the resource]." And I'm like, "You do?" Because they have no idea that I'm involved. And then to hear that is just really exciting. You know, making that information available means the world to me.*
>
> *(Valentina, Pos. 164-165)*

Combined under this register of valuing are the appreciation of working conditions, learning experiences, and emotional gratification. These ways of valuing highlight the individual perspectives of every person involved in the process, bringing together and aggregating data in these biodata resources. I want to highlight this aspect, as it brings forward the personal perspective that exists within and also outside of the professional positions everyone involved takes on. As all the practices explored below are performed by people, this aspect is important at every step. Still, biodata resources impact more than just individual people who interact with them.

5.2.3 Community Value

As central infrastructures of research communities and as collaborative spaces for contribution and exchange, biodata resources cater to groups. Their features and the doings of their maintainers are valued in this regard. Interviewees talk about resources as collaborative spaces, as driven by connection, and as not-so-human members of communities themselves:

> *It provides more of a sense of community within the [research field]. I think – I know, I'm not the only person that feels very strongly about the database and is very proud of it, even though we only […] provide these manual curations and use it, but we still feel very proud of it. I think we have a relatively close-knit research community, and I think that is a part of it. [The resource] is definitely a part of it.*
>
> *(Liv, Pos. 141)*

Value in terms of community is not limited to small, familiar research fields. To the contrary, resources can be valuable for their ability to bridge the social distances within or between fields. Bridging academic distances is then not just a matter of making data travel, but making researchers reachable across epistemic boundaries. When viewed from this perspective, biodata resources are seen as a driver or at least an enabler of collaboration.

> *If I'm using something here. I want to know more about it. I can find out who… You know, maybe I can collaborate with that scientist, because surely they'll have greater insight, and we can work together. So part of this is actually a sort of social thing – I think it should be a social thing, to allow people to connect and to collaborate.*
>
> *(Nico, Pos. 133)*

As will become clear later, such collaboration is not just the result of resources but also an important factor for the maintenance and development involved. In discussing integration and standardization across fields, resources are not just discussed as part of their field but as members of larger research

collaboration efforts. As such, community is an important register of valuing. It becomes important to highlight valuation and practice as embedded in larger social contexts.

5.2.4 Openness & Control

Operations of data aggregation and standardization, of making useful and usable, are inherently tied to notions of openness and accessibility. The biodata resources described by the interviewees are efforts to make the knowledge of large research communities available. At the same time, structure and consistency are similarly important, requiring guidelines, rules and often some level of control. Explaining important decisions, maintainers talk about balancing the two. This last register of valuing is one of ideals and of contrasts. Two extremes that become relevant again and again – when referencing data at large, when designing and enforcing contribution guidelines and in relation to managerial decisions. Inherently, the two need to be negotiated, as the advancement of one seems to coincide with the questioning of the other. Value in this case is discussed in regard to striking the right balance. Before exploring the underlying negotiations, I want to provide an overview of the different factors at play. Starting with the value of openness.

> *There are some people who aren't playing well in the system, and that's a shame, but because other people don't, it's not reciprocated. There are attempts to sort of fix that, but largely, people share their data, and that's where the value comes from.*
>
> *(Nico, Pos. 124)*

Openness means sharing the output of one's research with the research community and the world at large. Something that is not necessarily tied to resources, but can be facilitated by them. Their ability to do so becomes yet another factor contributing to the way they are valued. But openness is not just about facilitating the ideal of scientific distribution. More concretely, such an openness is valued for enabling the uninterrupted travel of data. It is a function of accessibility. The community-curated knowledge bases at the center of this thesis further extend this idea. Here, openness is not just important when talking about accumulated data; it also shapes possibilities of contribution.

> *This is the absolute best way to do curation, where the stuff that is being curated is on this website. Everyone can see it. Everyone can see as it's being curated, and everyone can make a suggestion. But either as informal suggestion, or they can actually make the fix and then say, "Hey, please just accept this as fixed" […]. So there are people who are part of the project, and there is everybody who is external. And that's a pretty great feeling.*
>
> *(Louis, Pos. 28)*

Louis highlights the way openness is valued. Here, everyone can contribute to curation and even provide corrections. However, the quote also makes palpable the connection between openness and control. The final decision on a submitted fix is reserved for internal reviewers. Louis talks about openness, but at the same time, he mentions the hierarchies embedded in the process. For him, it

needs both the external contributors who are free to suggest changes and the internal reviewers who hold the authority to check what is submitted. All of the maintainers interviewed described their resources through this coexistence of openness and control. What is seen as the right balance once again differs from context to context, between positions and those filling them.

This register will also play a larger role in the following chapter, not just in the relationship between maintainers and users, but also regarding institutional and external stakeholders. Seeing biodata resources through this lens provides a new perspective that is quite different from the primary and even many of the secondary registers of valuing. Once again, specific features are taken into account, and once again from a specific standpoint. While openness seems to be desired by many, this openness needs to be enabled and upheld in a way that leaves intact the resources' long-term goals and achievements.

It becomes clear that these ways of valuing are not necessarily tied to only the data, only the aggregate or only the resource. Primary registers of valuing are tied more to these epistemic properties of biodata resources, considering knowledge as it is made movable and presented through a number of factors. Similarly, usefulness and usability are closely tied to the relevance of aggregated data and the ability to access and interact with it. Differently, however, secondary registers of valuing engage with epistemics only peripherally. Biodata resources' usefulness in relation to academic metrics determines the perception of databases and often coincides with the motivation of community contribution. However, personal benefits go beyond these metrics, ranging from career development, through learning opportunities and to a feeling of pride and accomplishment. It becomes clear how different stakeholders make sense of databases in relation to themselves and to others. As such, contributions to research communities and the importance of collaboration to achieve shared goals are omnipresent in the discussion of resources as well. Resources are shared digital infrastructures aggregating dispersed knowledge, making openness and collectivity central features, though not without guidelines and a well-balanced level of control. In every practice, these perspectives are enacted, while themselves deeply connected and situational.

As analytical lenses, these registers of valuing will become important in the next chapter. To unfold their analytical power, ways of valuing need to be combined, highlighting how they are at times in conflict and need to be made compatible through negotiation. For example, usefulness to one user or even usefulness to a whole field might not be the same as for another. There are many ways to value biodata resources that influence what is done and how. Nonetheless, beyond these registers, valuing depends on multiple situational factors, including positions, rules and valuation infrastructures. The registers identified up to this point will help guide the analysis going forward, making visible this endless push and pull of different ways of valuing. They extend the theoretical concepts introduced above, adding additional vocabulary to the analysis of valuing and valuation constellations. In the following chapter, these concepts will be employed collectively to better understand how value is enacted through practice in the context of biodata resources. This will help answer the question of how data are valued.

6. Valuing in Practice

> *It starts at the very beginning. When the data are collected, the experiment's design… It's all captured, ideally in a systematic way. That's made available as part of the scientific record. The data is shared as openly as possible. They're integrated into larger sets. They're connected across multiple databases. People go and access them. They add their own data. They make that available as well. They interpret, they write scholarly literature that's connected into the system by further curation activities or by direct citation of data and the whole thing, cycles and cycles and cycles and adds, and continues to add value. So there's no one point of value addition. It's a continuous thing.*
>
> *(Nico, Pos. 84-85)*

I now return to the practices. Finally. Talking about valuing inherently comes with a need to look at how these values, the evaluation and the valorization, are embedded into doing. As data are handled and decisions about their journeys are made, their value is constantly assessed and enacted. As I laid out before, looking through the lens of valuation studies means looking at practices as enactment of valuation. Every practice inherently incorporates evaluation and valorizing. In this section, I take on this lens while looking out for differences in how data are valued at different stages and by different positions involved. Accordingly, this section is divided into three parts, looking at practices that serve different functions in and around resources. Each section will itself explore a number of practices, highlighting their valuing nature. Throughout, potential synergies, conflicts, and negotiations will be highlighted, making visible the integrative nature of these practices.

The first section explores what is often referred to as the data pipeline, starting at the center before moving outward. Here, the data are processed in standardized ways – selecting, curating and making available the knowledge that is so central to biodata resources. The practices highlighted here are often direct interactions with the data, shifting and changing them throughout. The data are made movable. Be it through professional curators, community members or automated systems, what is done in the data pipeline is done *to* and *with* data. Especially the primary registers of valuing are relevant here. But not every practice by users and even curators is done within this process. Not everything they do is part of the pipeline. Researchers, curators and users interact with the data pipeline from the outside, in publishing research, contributing time to curation efforts, and using the aggregated knowledge. Their actions are crucial to feeding the data pipeline, to making it sustainable, and to ensuring it is worthwhile. Consequently, more than just data is valued in these practices. Still, in this second set of practices, contributors are acting as extensions and outside beneficiaries of the resource's structures. The development and maintenance of these exact structures make up the third category of practices. They are performed by maintainers, peers, and other stakeholders and therefore, shape data directly at the individual level. Instead, this infrastructuring changes data processing as a whole, influencing the way that data are made movable and are valued in the pipeline systematically. In extension, these changes also influence the interactions that users and community curators have

with the resource, highlighting the impact of decisions made on this level. Here, all registers of valuing identified are at play.

6.1 Processing data

> *Curation is the process of organizing data into a structure that is consistent and understandable by the machine or whoever's going to have access to it next.*
>
> *(Nico, Pos. 88)*

Curation, or more generally, data processing in biodata resources, happens along a number of steps, not always neatly ordered, but often in sequence. Starting with the identification and selection of papers, the first decisions on what to include in the database are made. After prioritizing, papers are described and annotated in standardized and predefined ways, fitting them into the larger collection and making them ready for potential integration. Lastly, the integrated data are bundled and made consumable, highlighting specific entities and aspects, while making them explorable with visualization tools or other features of the website. The data are now movable, ready for re-use.

6.1.1 Selecting Data

The knowledge base maintained by Hannah selects papers regarding a certain model organism as they are published. Most papers are identified through an automatic system that scans a large repository for keywords. Once identified, the authors receive an email asking them to curate their publication. It is an established system that is so quick that some authors learn their publication is finally published through these emails. But not all papers are new, not all papers are caught, and, especially amongst the older papers, not every contribution is as relevant. This means that next to the automated system, the backlog and curation list are changed manually, based on community suggestions or whenever papers become relevant in current research discourses.

> *So that would be one thing that could be looked at if we could say, "right, these are all the papers that were published between 1990 and 2000, and of these, these are the ones that have been curated", but then again, I'm even – I'm even wary of that because we tend to curate papers when they're referred to in other papers that are newer because then we can tell: this has some information in that's been reported in this newer paper that we haven't yet gotten in the database and then we'll go back and do it.*
>
> *(Hannah, Pos. 31)*

While some of the curation of old papers is driven by the community and their requests, most of the curation follows the publications that have come out. The system is optimized for completeness in a way that deems the newer publications and those that have become relevant to the current academic discourse more valuable. Beyond the curation of new publications by community members, curation is done by a small team, which results in the rest of the backlog getting deprioritized and worked through slowly.

> *So [the new papers] either get done [by their authors] or they don't, but then the papers… we have a big historical backlog as well. So I think we have […] 6,000 papers that we think could be curated, and so far we've curated 4,000 of those. Now, of those 1000 have been curated by the community and checked by us, and the other 3000 have been curated by me and the other curators.*
>
> *(Hannah, Pos. 31)*

This system enables the research community around the resource to have a representation of current developments and relevant literature. It balances complete knowledge in terms of being up-to-date and in terms of relevance, prioritizing the usefulness for the community. Here, decisions are driven automatically by the digital infrastructure and by expert curators who are in contact with users. But not all resources follow this model. In other cases, users of all kinds play a more central role. The resources maintained by Jan and Valentina, respectively, provide a platform for users to curate what they deem important, regardless of timeliness or larger community needs.

> *The reason I started this thing, like I said, is to have a resource that people could go to to ingest the information.*
>
> *(Jan, Pos. 127)*

> *We found this person that was just adding data into [the resource], and it was high quality. We were accepting his changes a lot, and it was like, okay, this is awesome. […] And he was like: "It just captures everything I need. I can put it in there, I can find it whenever I need it." Basically, we found the user we were aiming for.*
>
> *(Valentina, Pos. 117-119)*

Here, users of different kinds ingest data whenever they want. Amongst them are researchers collecting the literature they read, entire organizations or institutions using the resource as their curation interface or protein developers who want to make visible their newest development. Regardless, the selection process is not built around a centralized idea of complete knowledge and usefulness, but around the valuation of research by each user or user group. This also means that curation is initiated from the community side, not sought by the resource. While all three resources aim for completeness and usefulness in their collections, the way those factors are defined differs. As a side effect, especially active users of the latter two resources can significantly shape the collection assembled. This shows the degree to which the valuation of the knowledge aggregated is dependent on perceptions of completeness and usefulness. What is apparent as well is how different positions are embedded in rules and infrastructures of valuation. While users of one resource have no direct influence on the selection process, users of another are central to the selection of research for curation. The registers of valuing are the same, yet their enactment is structured differently across resources, as rules and infrastructures differ.

6.1.2 Curating Data

While all of the resources considered in the previous section focus on research papers, other approaches are also thinkable. One such approach could, for example, not take papers as a whole; instead, the process could involve the active search of information across the literature.

> *My aim is more of an integrated, holistic model of eukaryotic biology rather than these independent pieces – units of information. It's just that the publication is the unit that we curate by. I mean, some databases may not go by papers; they may go by gene and just be looking for information about a gene.*
>
> *(Hannah, Pos. 23)*

Either way, the extraction of information is at the center of the curation, regardless of the unit curated. This curation can differ in a number of ways. First of all, papers can be just described, resulting in a list of papers that can be easily searched. Ideally, such a process would include standardized terms allowing the integration of this collection into larger contexts. Beyond the description and integration, extracted knowledge can also be tied to larger models of biology. This is often the case in model databases, like the one maintained by Hannah. The resulting "holistic model of eukaryotic biology" is built on these papers, while moving beyond their fractured distribution of knowledge. Here, the idea of completeness is evoked, but this time in relation to a structured "holistic model". To make data movable in this way, certain practices are central to all curation.

> *You read just the results and look for… You know, read the methods first, figure out what they used, then you would curate that to: okay, here are the different organisms, different mice they used. Here's precisely what's in that genotype for that mouse. And then here's all the phenotype data, and oh, this is a new phenotype, and we don't really have the right term. And so then you go make a new term in the ontology. And sort of cycle around.*
>
> *(Sophie, Pos. 58)*

Within these processes, it is often clearly defined what information can and can not be captured. The case of "new phenotypes" is an ontological issue, resulting, in this case, in the extension of the ontology – a procedure that itself is usually pre-structured to allow a controlled extension of the terminology. Based on these rules, only some data are seen and enacted as valuable. Other pieces of information might not be as easily included as the fields provided to the curators only allow for certain common or relevant types of information. Only what can be valued through the infrastructure becomes a part of the aggregate.

> *We are very often […] not the union, but the intersection of data that is commonly reported in the literature about properties of these things. And that's a bit of a bummer because the union would be much more interesting. It's just that if we did the union, you would go to a given page, and it would just have a single little bit of data that was only measured in this one paper, and it doesn't have any of the other.*
>
> *(Jan, Pos. 98)*

This question of what to include and what not to include connects to a number of ways that data aggregates are valued. Detailed and holistic approaches might improve the completeness of knowledge, but a consistent structure requires a limit to the extent to which new data formats and kinds of information are included. This negotiation, in turn, impacts the availability of relevant data on the one side and the usability of less or more structured and consistent aggregates on the other. In addition, correctness and trust are also important.

> *Anybody can curate. So, anybody can add evidence. Anybody can change evidence in [the resource], but we have an expert moderator team that has to approve those changes.*
>
> *(Valentina, Pos. 110-111)*

Such a two-level approach is present in all of the community-curated resources maintained by the interviewees. In some cases, moderation means collaborating with community-curators to find the best curation; in others, changes are just accepted or rejected. The valuative work here is a two-step process including multiple positions, that is deeply embedded and stabilized. While, as mentioned before, these curation processes are designed to create consistent and reliable data, they are not completely fixed in their design. By including new tools, like the automations, by applying changes in the ontological vocabulary or by modifying the types of data captured more generally, curation practices can shift over time. Such a shift might require data that has passed the pipeline before to pass it again.

> *I would say there is a need [to re-curate], but it didn't necessarily get done. […] But also, that reflects the state of the literature as well. […] There are just less people studying [the specific] properties.*
>
> *(Jan, Pos. 94)*

In this case, the missing re-curation is the result of missing data, leading some fields to remain empty. While the knowledge aggregated might remain incomplete, it still represents the field somewhat completely. In this case, the value of the aggregated data is connected to the representation of the current state of research. Complete knowledge is tied to what is available, and less so to the underlying biology. In other situations, re-curation is more pressing. If the information provided is identified as wrong by users, maintainers or cooperation partners, re-curation becomes a priority to

ensure the knowledge provided is correct. If the rules were not correctly applied or have been revised, curation needs to be repeated to render data and their value movable once again.

> *Somebody from another database emailed me today that we used the wrong enzymes… the GO term for an enzyme, and it was obvious when I saw it. It was one of mine, actually. And I misinterpreted what the GO function meant, and they were very similar.*
>
> *(Hannah, Pos. 46)*

Correctness is not just the result of a fitting structure of the pipeline, or the availability of correct ontology terms and curators able to use them. As curation is done by people, their attention to detail and the work environment they work in similarly impact curation. Looking back to past experiences with supervising curation efforts, Louis recounts:

> *[Other teams] did work to get the numbers done, not to get good content. And so that was the wrong perspective, in my opinion, because the students weren't doing a good job, and the content they created wasn't useful. […] On the other side, the work itself… […] We would go through curation together, and they would find, you know, corner cases everywhere. Like, here's something we're pretty sure fits into what we're doing, but we can't represent it using the curation system that you've devised, right?*
>
> *(Louis, Pos. 66-68)*

Regardless of the quality, the result of curatorial efforts is a dataset that represents either a subset of the literature or the aggregated knowledge about a certain entity. This is achieved either through paid curators, community members or automated systems. In the process of curation, completeness and correctness are enacted by the rules and infrastructures as much as by the people making data movable. The exact factors that make data correct and complete are closely tied to the potential re-use, making the *usefulness for…* yet another important register considered in the process. The resulting aggregate, its underlying structure, and format ideally enable further use, potentially providing new emergent knowledge, potentially enabling new methods, and definitely providing some overview of what is known.

6.1.3 Communicating Data Aggregates

Aggregating and curating data is not all that happens throughout the data pipeline. Knowledge at this point is already brought together and, to some extent, decontextualized, but it is not made usable through interfaces and tools. Data are still in their database format, not communicated and recontextualized. They exist in bulk, not bundled into sensible and relevant packages. While many of the resources described by the interviewees started out as Excel sheets or JSON files[4] containing only the data, their interfaces make them usable to a wider audience. For one, they are available across locations and communities in a way that even a shared spreadsheet could not. Further, their interfaces

[4] JSON is a file format that allows the easy storage and transport of structured knowledge between programs. Due to its open structure, these files are easy to interact with as well as (somewhat) readable for humans.

allow interactions with the data in ways that don't require extensive knowledge of data handling. Most importantly, however, these interfaces provide descriptions of the data collected and of the metrics used to describe them. Such information is not only informative, but might also highlight shortcomings of the data provided.

> *Perhaps the best case of this is called photobleaching. […] The measurement of it is so critically sensitive to every parameter that the experimenter used when they measured this thing that it is almost impossible to make anything out of what this person measured and that person measured. If one person measures the photobleaching of 10 proteins on their microscope with their light source, with the same exact settings, that's kind of interesting. But if 10 different labs measure, they're gonna probably get different relative amounts.*
>
> *(Jan, Pos. 109 & 112-113)*

Communicating data is not just about the context of creation and the definition of measurements. Communicating data correctly makes the most relevant data points the most prominent. Between the context of data creation and data use, the communication of data at the end of the pipeline is the place where movability is assembled. By combining everything that needs to be known about the data creation and that is relevant for users, the resources strive to make the data usable. Yet as usability is valued differently across user groups, this last section of the data pipeline is just as contested as the ones prior. Talking about a meeting with the Scientific Advisory Board of the resource Hannah maintains, she describes this as a challenge, given the diverse research backgrounds of potential users:

> *And there's a little bit of conflict about which information should be given the highest priority on the page, because some people are only interested in one thing, and some people are only interested in something else, and we can't give every single piece of data.*
>
> *(Hannah, Pos. 42)*

Presenting information in such a way is not just about prioritizing the most important aspects. Often, resources also communicate curation to improve trust and give insight into the data pipeline. This includes the provenance data highlighting where information comes from, but also descriptions of the curation processes. By communicating data, resources not only allow users to value data but also value the processing that was done to them. Users learn how data were created, how they entered the resources and how they were processed, bundled and visualized. This communication is aimed at building trust and providing user the opportunity to compare their own perceptions of valuable data against those of the resource. To provide usefulness beyond the information collected, additional tools are added, allowing users to visualize and analyze curated data. One such tool prominently featured throughout the interviews is AlphaFold, as mentioned in the section on useful knowledge. But proprietary tools, tied to each resource, are just as important, empowering users to interact with data

in new ways. Beyond the interface, resources not only prioritize certain features on a page. Data are also bundled into larger collections as well.

> *So we've just started to building [data collections]. So, on the front page, here's an example: […] each of these genes and chemicals linked together, by inputs and outputs. We've done quite a few of these. […] This is how we're pulling all the information together.*
>
> *(Hannah, Pos. 48)*

Integrating selected data in such a way means that a resource can provide a detailed picture of a section of biological knowledge. This bundle is optimized for usefulness, and additional papers are, at times, curated to fill potential gaps. While broad in scope, this selection is aimed at providing complete topical knowledge by deprioritizing some data and further integrating others. Such curation is once again aimed at a certain use. Talking about the design of this part of the data pipeline, all registers of valuing are employed. Here, the knowledge, usefulness and usability value of resources come together. The result of this negotiation can look quite different. In the case of Jan, the approach is quite opposite to that of Hannah. On the website of his resource, data are also presented, however, additional effort is put into providing data in bulk through an API. Such an API is aimed at users extracting data in a computer-readable form, meaning that fields are not described, warning labels are omitted and that bundling is reserved for the users themselves. With those omissions, valuable information about the context of the data creation and curation might be lost, putting into question whether the data will be interpreted correctly. Coming back to the case of photobleaching mentioned above, Jan reflects:

> *I think I would probably aim to have a discussion with somebody [extracting data in bulk] and try to educate them about cases in which noise… variability of the underlying data is likely to be higher than they expected it to be. But […] sometimes they get the data, walk away, and we never talk again. And so that conversation doesn't always happen.*
>
> *(Jan, Pos. 155-157)*

This communicated data – both in computer-readable form and on highly descriptive user interfaces – is what users interact with to extract the data and knowledge they need for their research or subsequent uses. With different use cases, expectations for this part of the data pipeline might differ. The different kinds of communication are valued differently across contexts. The relevance for research is not only a matter of selecting the right research to aggregate or of curating certain aspects with a relevant vocabulary, but also of bundling the aggregated data into a format that is both useful and usable. Through the data pipeline, data are constantly valued in regard to the features of the aggregated knowledge, the usefulness for academic and downstream use and the usability provided for certain kinds of interactions. It is striking how, within descriptions of the data pipeline, the focus lies exclusively on these primary registers of valuing.

Designed by maintainers and enacted by a number of human and non-human actors, these socio-technical infrastructures are built to process data. As such, they also stabilize certain ways of valuing. They do so in highly structured ways, with predefined selection, standardized curation, and technically embedded communication. Yet these processes are not automatic, in some cases allowing contributors to suggest the research taken into the system, while relying on manual curation to ensure relevance and correctness for subsequent use. As such, interaction with the resource is not just limited to the end of the data pipeline. Across positions, for authors providing their research, curators committing their time, and for users of bundled and visualized data interaction might look quite different.

6.2 Interacting with Data Pipelines

> *And so I think there's some people who like to just make sure it's complete, and if they get to be the one who puts this new protein in there, they feel a little goodness about that, or something like that. So I think it depends on who's submitting.*
>
> *(Jan, Pos. 57)*

From selection, through curation, to use, people interact with the resource and its data processes constantly. As far as one can see them as distinct, each researcher, curator, and user brings different needs and motivations, looking to contribute to the resource, to benefit or both. As resources can be quite embedded in research communities, these positions might be united in one person or spread across many. Either way, they come with types of interaction that I want to highlight in this section, enacting ways of valuing that are about more than just the data and knowledge aggregated. The primary registers of valuing, so important in the previous section, will still shape these interactions, but the secondary registers play a role as well. Here, non-epistemic, personal and communal benefits of different kinds, and ideals of openness and control are negotiated. In this section, I want to describe how values are enacted when researchers publish their research, thereby building the foundation for data aggregation, when curators of different kinds move along data throughout the processing pipeline and when users at the end of it all provide purpose to resources through their use.

6.2.1 Publishing Research

The practice of publishing research is in and of itself independent from biodata resources or aggregation efforts at large. Depending on the field or the aim of the researcher, resources might not be considered at all at this stage. Nonetheless, researchers are often aware of the path their output takes. The outputs of researchers, labs and institutes provide the foundation for any of the work described in this thesis. Without data, there is no data aggregation. The act of publication is closely related to the aggregation and integration by databases. In this section, I want to explore the entanglement of research publishing and biodata resources.

Depending on the selection methods of the resource, researchers wanting to have their publications represented need to suggest or even curate their output themselves. Only then can their research move and be useful to others.

> *We try to make sure that we notify the relevant databases to have the data curated on there, but they're very different processes […]. So, for instance, [this one resource], you essentially just send them the paper and write a blurb about what you think needs to be annotated, and they do it themselves.*
>
> *(Liv, Pos. 28-30)*

Research that is not actively suggested or is selected by the resource itself still finds its way. As the published paper already comes with metadata describing its content through keywords, abstracts, and further descriptions, the needed prerequisites for automatic selection and curation are provided. In this step of adding metadata to papers, researchers are not the only ones involved, as publishers also aim to make the publications travel as far as possible. Especially in large-scale repositories like the one Bejan is working on, this information might be the only thing considered by the system.

> *The data we get from third parties, we don't really… You know, what we get from PubMed and PubMed Central. That's what it is. In fact, they get it from the publishers, right? And the quality of what the publishers send them can be quite patchy, but we don't interfere with that. If the publisher sends us bad metadata, we don't fix it. That's kind of a matter of policy […]. You can find all kinds of mistakes in the metadata if you look for it.*
>
> *(Bejan, Pos. 26)*

Publishers, in this case, provide metadata but do so independently of the resources that will distribute their publications to potential re-users. In their position, researchers and publishers only have limited agency to determine how the value of data is transported through the system, while automated processes make the data move.

Around knowledge bases, like the one maintained by Jan, efforts are more coordinated. As described before, his resource is often used by protein developers aiming to distribute their new proteins to potential users in the research community. Seeing the value of potential dissemination, their publications are written in a way that is compatible with future aggregation, even including the accession number[5] provided by the resource. But this integration also comes with challenges for the researchers, as accession numbers are only provided to entries with a DOI attached, and the DOI is only available after the publication. The values stabilized by the system are targeted towards correct and proven information in a way that is in conflict with the researcher's goals.

[5] An accession number is a unique identifier provided by the database to every protein. It allows clear and long-term distinction between entities even as human-readable names change.

> *This is how new data comes in. So somebody goes to submit a protein, and the two mandatory fields are a name and a reference [requiring a DOI]. And this is interesting because it makes a little bit of a chicken and the egg thing. Where, if a protein developer wanted to submit their data so that they could have it in their publication, like, what is the accession number? That is a case where they need to reach out to me directly. I will manually enter it with a sort of pending entry.*
>
> *(Jan, Pos. 42-43)*

Publication, in this case, is closely tied to the curation, going so far that the curation process itself needs to be extended slightly to allow the alignment of publication and database entry. The stabilized processes are extended with circumventive measures that negotiate the conflicting ways of valuing data and data work. In other cases, not only the description of the research published, but also the selection of the topic is tied to the resource.

For Liv, a user, research provider and community curator, these different ways of interaction are integrated in multiple ways.

> *They have made this list of all the proteins that are uncharacterized […], and so that's one of my kind of interests… Characterizing these uncharacterized proteins.*
>
> *(Liv, Pos. 34)*

Here, research gaps are identified by the resource, in turn leading to research that complements the aggregate. Research publication and aggregation are integrated tightly. The valuations between the resource and Liv as a user and researcher are aligned. She sees the aggregate and its communication as useful and, in turn, produces research that perfectly complements the resource's development towards *complete* data. But as Liv not only uses the resource to plan research and has her publications represented in the database, she also commits time to support the curation process itself.

6.2.2 Contributing to Curation

Data do not move through the pipeline themselves. As automated systems fall short in terms of quality and reliability, human curators need to support the data's travel. Liv is a community curator for a resource where researchers curate their own publications. This work is not mandatory, but in this case, it is the only way that her research can make it into the aggregate. Due to her familiarity with the resource, she has taken over the curation for the lab she is a part of.

> *I mean, I genuinely actually enjoy curating the papers. And so now [the resource maintainers] just send [the curation invitations] all directly to me. Even if I'm just like a middle author or whatever, [they] know that I will happily do it.*
>
> *(Liv, Pos. 22)*

In the quote, she describes curating as something she enjoys – providing gratification, closure for finished projects and a sense of pride. She values the resources not only for their knowledge,

usefulness and distributive power but also ties her engagement to personal benefits she derives. For both her and Luka, contributing curation to resources important to their research fields is tied to feelings of giving back, providing more than personal enjoyment, but also a sense of community. In their personal decisions, the secondary registers of valuing play a substantial role.

> *I think the main motivation is to share what we have done so that it's out there and it's available to the […] community. And also, I think that's the biggest reason why I do the curation, to just share it with the community, because […] I benefit from the information that's out there, that other people put out there. So, I want to do the same thing back.*
>
> *(Luka, Pos. 51)*

In the interviews with maintainers, the motivation of researchers curating their own publications is often related to a potential increase in visibility and citations. While they are not sure about the actual effect, they portray these non-epistemic benefits as part of an exchange, providing reach for the time and effort invested. In their accounts, they weigh one way of valuing against another. Luka, on the other hand, talks about contributing not only for her own benefit but for the longevity of the resource. She values the usefulness and, less so, the potential non-epistemic benefits she might gain.

> *I believe that curation – like, having the database curated – it's pretty important for the funding part, right? […] It's such a useful resource to us that it would really be a shame to, you know, have it not be funded and then not be there for us.*
>
> *(Luka, Pos. 132-133)*

Outside of personal motivations, curation can also come in more organized forms. While some resources employ curators as staff, others benefit from organizational partners using their resources as a tool to store and analyze relevant research in systematic ways. Regardless of the motivation, curation is a tedious and time-consuming process as each field needs to be filled diligently to ensure the correctness and standardized nature of the aggregated knowledge. This is where different perceptions of correctness and completeness might influence the practice of data processing. While the data pipeline prescribes a certain format, curators might not agree with the level of detail aimed for.

> *There's something that's specific to… You have to say "mild," or "high," or "low" in some respects of the curation. It's too much. Like, it's too detailed. And I'm like, "Okay, I'm not gonna do this", and I just skip that part altogether. […] To me, you can just look at the picture and judge it yourself.*
>
> *(Luka, Pos. 37 & 41)*

While still moving data through the pipeline, a part of the curation is omitted, as its value for the knowledge is questioned. Room for such considerations can be built into the system, with curation interfaces containing both mandatory and optional fields. Nonetheless, the result is a structured

dataset with missing values, highlighting the impact that curators have on the final aggregate. But they shape more than the level of detail. Even before the curation, contributors can influence what is and is not aggregated. In processes that put the responsibility of selecting relevant papers on the curators, their choices influence the shape of the collection at large. Valuation of research is not fully stabilized in the infrastructure, as some valuative agency is placed in the role of curators.

> *You see that they have patterns of curation that are all focused on, you know, myeloid neoplasms or brain tumors, and they focus on that area, and maybe they focus on a specific gene […]. So it kind of really depends on what their domain focus is. Most people that enter things […] have that kind of focus. I would say that one of the exceptions are those [expert panels] where they may have been funneled to be a curator and part of what they're trying to get out of it is actually to learn.*
>
> *(Valentina, Pos. 135)*

Especially in resources like the one Valentina works on, the curators' research interests can impact the collection. As resources serve as a systematic storage of read material, community members mostly curate for their own use. The ideal structure of data is embedded in the curation system, but definitions of completeness remain open for curators to decide. The other use of the resource mentioned in the quote is education. Either facilitated by educational institutions or as a side effect, curation brings familiarity with relevant ontologies and trains the extraction of information from research papers. For people trying to gain familiarity and for educators, the resources and their structured data work have an educational value, providing personal benefits to those who make data move. The applicability of ontological terms outside of resources is not perceived equally by everyone interviewed, but the educational benefits are generally regarded as positive whenever they were part of the interviews.

In curation, all the registers of valuing are interacting. In close interaction with the data, knowledge aggregates are produced according to both personal positions and the rules and infrastructures provided by the resource. Usefulness is defined differently across platforms, but in any case, the decisions of what are and what are not deemed useful data are enacted by curators. This work is done with the resource, but it is also shaped by personal perceptions that go beyond knowledge, usefulness and usability. Motivations to curate range from non-epistemic rewards to personal benefits, as curators invest time to advance their careers or serve their communities. Along the way, they negotiate different ways of valuing, at times pushing against the positions they are assigned and the rules and infrastructures that shape them. While the pipeline itself stabilizes more or less clear ideals for the data aggregate (see the above section 6.1), the human enablers of this process enact a more-than-data-focused perspective.

6.2.3 Using Data Aggregates

The largest group to interact with biodata resources is users. Researchers and research consumers from different fields and with different interests visit resource websites, search, interact and analyze

data, extracting whatever they deem relevant either in small chunks or as a complete dataset. For them, the resources are streamlining their work.

> *[Without the resource] we would probably be less aware of what's published. And I think as much as we try and have our hands in all the research and all of the papers that are published in our field, I think it would lead us to doing more things than we would need to do.*
>
> *(Liv, Pos. 148)*

For Liv, the resource provides accessible insight into the current research landscape, therefore making redundant research less likely. Beyond scoping the field, the visualization tools and analytical features enable a deeper understanding of the knowledge aggregated. Similarly, Luka values the usability and usefulness of the knowledge aggregated.

> *You go through the entire page and you see, okay, maybe this gene interacts with this, this gene interacts with that, and you might have an idea, you know, "I'm interested in this gene because it's affecting X process, right?" Then you go on the website, and it's like, well, it's also been implicated with that process, and perhaps there is some way to connect that process to the process that you're interested in. So, again, highly valuable because all of that is within one place.*
>
> *(Luka, Pos. 121-122)*

This exploration is not just about finding knowledge. It is also about finding new and helpful ways to interact with the information available. Both in terms of data and in terms of the tools provided, each community, even each researcher, might have different needs. For users and for maintainers talking about usage, the knowledge, usefulness, and usability seem to be important indicators of the value of a resource. How different these primary registers of valuing can be enacted becomes visible in the example of visualization tools.

> *There is a specific way to excite [fluorescent proteins] that is a little unusual. And it's a niche application that, within that community, it's all they care about, right? And so that community did sort of early on say: "Can we get two photon spectra? Can we get two photon?"*
>
> *(Jan, Pos. 90-91)*

Other than the previous types of interactions – researchers publishing and curators contributing – the impact of usage on the content is limited. Inside the data pipeline, different positions, curation rules, and data infrastructures shape the movement of data, but for users, valuation can only happen in relation to the final aggregate. Their position at most comes with the possibility to request changes, while their usage or non-usage is reflected in impression numbers. Still, the presence of users is what makes the resources lively. Users' numbers are employed to describe the relevance of the platform to funders and other stakeholders, while maintainers go to great lengths to understand who those users

are, what they do and what they are looking for. While not directly involved in the curation and maintenance practices, users are implicitly involved throughout most of the decisions. Their presence and usage patterns act as a proxy for the usefulness of a resource.

> *We talk to people about what they want, we run feedback surveys from users regularly. We do. So, I mean, it's not me, but we run interviews as well, like… and observing people using the [the resource]. […] What we're doing is, I would say, pretty far outside the norm for an academic group in terms of developing a product.*
>
> *(Bejan, Pos. 17)*

This is especially important as many resources started out as very local tools, initially serving their maintainers or a small community surrounding them. As userbases grow and change, more ways of valuing intersect, requiring maintainers to consider more than just their own perspectives. Yet, while the role of users might be rather passive, this does not relieve them from all responsibility. As consumers of aggregated data, they are missing a lot of the context available to researchers and curators. As data travel away from their context of origin, this context has the potential to get lost. The distance becomes too great. While structured data might provide new context and integration, the usage of such aggregated data requires deeper and extended engagement with the underlying literature.

> *You know, you sort of have to trust that other people are doing thorough research and that they are doing whatever needs to be done to fact-check. But there's no way to know that.*
>
> *(Luka, Pos. 104)*

There might be a conflict here between users' trust in the aggregated data and trust in the users to interpret it accordingly. Correctness, while valued by all the interviewees, needs to be enacted both within the resource and outside, both in the curation and in use. Consequently, trust in knowledge-use becomes an additional consideration, adding to the valuation of trust in the knowledge itself.

Across all of these interactive practices, valuing leaps beyond the primary registers, beyond knowledge, usefulness and usability. Throughout secondary registers of valuing are enacted. In providing research, contributing to curation and using aggregates, values are combined and negotiated, at times pushing against structures and stabilizations. Still, researchers, curators and users only have limited power to change the resource beyond the positions allocated to them. Their impact is not to be diminished: Researchers provide the academic corpus on which resources are built, developing their fields and the resources that represent them in certain directions. Within the structures of the data pipeline, curators define what is relevant both on larger and smaller scales, ensuring usefulness. Their contribution is key to ensuring the value of knowledge as they extract, standardize and integrate what is interesting to them and ideally also to users. The latter, the users, are mainly tasked with making all of that work worthwhile, realizing the academic and societal value of research and its aggregates. Not just they, but all of those interacting with biodata resources hold a

level of responsibility, ensuring that they advance research as a whole. They hold responsibility but only limited control.

6.3 Infrastructuring Biodata Resources

> *We keep looking at new tools and new things to support, and you know, where can we safely automate? Where can we cut back and have some of the repetitive tasks get handed over to algorithms, and where do we really need human intervention? And we definitely do not have the answers for that yet. We keep working on it.*
>
> *(Sophie, Pos. 66-67)*

Biodata resources, when described in brief, are digital infrastructures tasked with collecting, integrating, and mobilizing biological research data to enable further research, providing data affordances, and doing so in a way that saves time and effort for the community they are serving. But these resources do not just exist. By the hand of their originators and maintainers, resources are designed and kept running, constantly subject to changes and refinements. They are optimized for their use, for integration and for longevity, but in the analysis so far, it becomes clear that such ideals might look quite different. What a "valuable" resource is depends on who uses it, for what it is used, and in what contexts it is supposed to be integrated. Throughout this section, questions of longevity will be discussed as well, as a central concern for maintainers. While those maintainers often work alone or in small teams, their struggles are shared by many others. Across projects, those involved in upkeeping and further developing resources often collaborate, especially on shared rules and infrastructures like the ontologies that describe their data. This development of standards, not just ontologically but also technically, is the basis for the integration of knowledge across resources. In doing so, maintainers and their peers create a context for data integration and mobility, both on the level of single websites and on a larger networked level. But maintainers and their resources are part of outside contexts as well. Funders, institutions and regulators provide a backdrop – and financial foundation – for the work done in and on the data pipelines. However, if present, this support comes with additional ideas of what *good* data and *good* data work look like. This can lead to mismatched imaginations of what, for example, a sustainable resource should be.

Throughout the remainder of this chapter, valuation will become visible as contested and in need of negotiation. Going forward, I lay out the practices identified across maintenance, coordination and context creation, highlighting once again, how they are enacting value across different registers. Throughout, I identify maintainers as key figures who relentlessly keep working on their resources. They hold the agency to create rules and build infrastructures, taking on a lot of the stabilization effort required.

6.3.1 Maintaining Data Pipelines

All databases started at some point, but not all started in the same way. While big research repositories are at times the result of centralized efforts to collect academic knowledge, many others, like the ones

prominently featured in this thesis, originated rather locally. In these first versions, usefulness was not always envisioned as something tied to a research community. For example, in the case of Louis, the resource started as a local tool, solving problems related to his personal research.

> *I needed to standardize data, and I coded up this dictionary of rewrites, and I thought. Wow, you know this is getting really long inside my code. Maybe this would be better inside a JSON file, and maybe adding a little bit of context to that JSON file, like, it'd be good to have the titles to go along with each of these things.*
>
> *(Louis, Pos. 40)*

The aim of this resource is to translate between ontologies. Initially not designed to be easily expandable, the structure changed throughout the process, developing from hard-coded scripts to a dynamic processor of ontology translations. This enabled moving all the translations into a dedicated JSON file, away from the underlying program processing the information. The switch meant that putting in new translations was no longer a task of coding, but now just a task of curation, creating a tool that could be useful to more than one user. However, with the publication, ideals of usefulness and valuable knowledge needed to shift as well. The value of the resource was no longer just individually enacted.

> *This is probably the key; how do you go from "this is something that was useful for my workflow" to "this is something that could be useful for other people's workflows in general" and along the same lines, "this has enough data curated in it to support what I'm doing" versus "this is complete and this covers everybody and everything". Of course, covering everybody and everything is a lot of work, and it's sort of an ongoing thing.*
>
> *(Louis, Pos. 52)*

While also initially local, other resources started with the explicit goal of providing a tool to larger groups. These resources were aimed from the start at providing a collective inventory of useful and valuable knowledge, at compiling overviews of what is known in fields that rapidly generate data, or at providing an infrastructure to systematize knowledge through the hands of researchers across fields. Still, their earliest version often looked nothing like the current resource, as their developers tested data structures in Excel sheets and programmed rudimentary user interfaces to enable the first curation and usage. Here, different ways of enacting value were tested and priorities negotiated before stabilizing them in rules and infrastructures. Throughout, each person or group working on a resource made choices about the ways that data travel through selection, curation and communication. In the beginning, these choices are based on the maintainers' knowledge of the field or on conversations with peers and potential users. However, to make a resource that becomes useful to its users, many things need to be considered.

While not working on a small and new resource, Bejan describes the struggles of early development. In designing a tool that could help curators within his organization, he hopes to capture their needs.

> *I don't know how much workflow change there has been so far based on my tool, but I guess my aim is to fit into the workflow. I don't want to change their workflow necessarily too much, you know? Because they have a way of working that… If their workflow is going to change, I think it would be through my tool being useful rather than me forcing them to a particular way of working.*
>
> *(Bejan, Pos. 48)*

As a data scientist, Bejan develops these tools himself, and in his team, there is additional staff to program and maintain the purely technical side of the resource. In most resources, however, such coding expertise is not available, and changes to the technical infrastructure are delegated to outside coders whenever needed and fiscally possible. In this context, technical changes become a calculated measure, undertaken only when necessary and with foresight. As the capacity to perform technical development and maintenance is limited, the task of gathering, synthesizing and implementing user needs becomes an important feature of the work. For Bejan, this process is one of talking to curators within his organization to program ever more fitting tools, but for maintainers of smaller resources with limited programming budgets, users and curators are not as easily integrated in the development process. Their ways of valuing data and data resources determine later success, making it crucial for maintainers to sense and correctly integrate how and with what goals their resources are used now and in the future.

> *When we do our survey, a lot of the PIs say: "people in my lab… everybody spends at least one to two hours a day on [the resource]." And I'm like, "I'd love to know what they were doing." Kind of, it's quite a long time. It would be quite nice to see exactly which bits of information they find most useful, but I guess it's different for every researcher.*
>
> *(Hannah, Pos. 42)*

As a maintainer, Hannah is central to making the resource valuable to the users, yet at the same time, what this value might be is opaque to her. Not all changes need to be anticipated in this way. As users approach the limits of what a resource can offer, some reach out directly to maintainers. Asking for certain changes to be implemented, the differences between the values embedded into the resource and the needs of users are communicated explicitly. These requests provide insight into the exact ways that users value the resource, but they come with an additional challenge. Balancing the needs of the individual and of all other users, maintainers want to ensure that their resource is not developing too fast in too many directions. Individual and short-term needs are negotiated with longevity and universal value.

> *We do get requests from groups, and it's like, "Well, you're not active. Like, why do you need this change? And what do you need it for?" And it's our convenience versus this will actually improve clinical processes or a research-specific research question.*
>
> *(Valentina, Pos. 85)*

Such considerations around longevity and stable structures on the one side, and usefulness and potential short-term benefit on the other, are not just limited to feature requests. In designing rules and infrastructures, maintainers negotiate values constantly, stabilizing them for future curation and use. In maintaining the data selection and curation, maintainers need to balance making corrections and building an aggregate structure that lasts. While keeping in mind every part of the data pipeline, small changes are weighted against larger structural developments.

> *Our original, very sort of hierarchical model… We literally blew it up and completely redid the entire website. […] We launched a V2, and literally, we retired our version one website and the entire database that went under it. So we converted everything to version 2 to allow for basically many-to-many relationships.*
>
> *(Valentina, Pos. 58-59)*

In this case, restructuring the database allows for a better representation of biology and the research describing it. While the content in the first version might not have been wrong, it showed only a partial picture. For the maintainers, this partial picture no longer represented the completeness they were aiming for. The change is a matter of achieving a collection of complete knowledge that is structured in a representative way. Changing the structure, however, might disrupt the work of those relying on the platform – a consequence that needs to be weighed up. In a different interview, Jan talks about not changing certain aspects to uphold current structures and systems for the users.

> *There are a couple of things I'd like to change, but I also have an API, and so I can't change it easily without a deprecation and a versioning system. So, I can't break any usage of people who are programmatically accessing the data. I can't just change the fields.*
>
> *(Jan, Pos. 74)*

Technical overhaul of the database often goes hand in hand with an overhaul of the way that data are described within them. The changes described by Valentina and Jan might therefore come with ontological changes as well. This would not only improve usability but also the structure of the knowledge as well. But changing ontologies is similarly challenging. While adding terms may be a standard procedure, creating sub-terms, changing relations, or even obsoleting items may require additional caution. Here, users are not the only group that needs to be considered; curators, filling the data structure, are part of the consideration as well. If done right, such changes improve their experience, as in the case of Luka. As a community-curator, her efforts are not paid for by the resources and rely on her motivation. When maintainers reconfigure curation rules and interfaces, the needs of curators might play an important role.

> *That's also something that just gets easier over time, and they have made changes to the process itself, which makes [curation] easier. So it's not as tedious as it used to be.*
>
> *(Luka, Pos. 147)*

Maintainers do not always decide in favor of curators when balancing the usability of curatorial interfaces with the completeness of the aggregated data. As additional information is sought, complexity and the needed work might increase. If, for example, a new field is added, new entries might require more time, while additional work is needed to complete entries of the past. This results in a more valuable collection, while at the same time, more strain is put on volunteers. Such extended curation processes also require additional training for those moving data through them. Providing education to curators and communicating data structures and vocabularies is another important task of maintainers. Here, they highlight not only what rules and infrastructures there are, but often also their value. The communication with those who interact with the resource is not just about providing curators and users with the required skills to manage the digital tools available. It can also be a crucial task in building a vibrant and motivated community that is aligned with the values embedded into the systems and the data aggregated. From a user-curator perspective:

> *And one thing that's helpful is that if you have questions or if there's something that… It's not coming up in the terms that are already there. You can always contact [the maintainer] or somebody at the […] help desk, and they're extremely, extremely helpful and very responsive. So again, that makes it, you know, it makes it fun.*
>
> *(Luka, Pos. 31-32)*

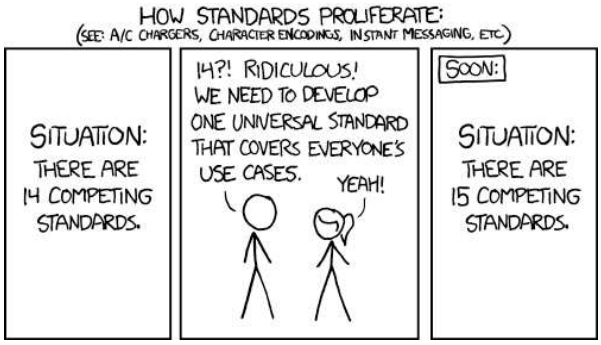
Next to building and maintaining their resources, maintainers build connections and communities that form around their collections. In this position, they build up a shared understanding of how to handle and therefore value data correctly and responsibly, balancing between ways of valuing that are closely tied to the knowledge and the more personal needs and benefits of those interacting with the resource. In the maintenance work, they not only build a resource that they deem valuable, but negotiate the many different ways that others perceive and enact data, data work and data infrastructures as valuable. Stabilizing situational and long-term value, closing conflicts and implementing resolutions, maintainers are central figures in the valuation of data and data work. Researchers, curators and users, while certainly a core stakeholder for their consideration, are, however, not the only audience relevant. Maintainers aim to connect to many other stakeholders as well.

6.3.2 Collaborative Structures

> *You have to think of all these databases as an infrastructure. They're an infrastructure for science. So they're equivalent to a particle accelerator for high-energy physicists. They're equivalent to a telescope for the astronomers. Yet they don't have a single physical presence. And, […] the infrastructure grows organically.*
>
> *(Nico, Pos. 43-44)*

The decentralized nature of biodata resources is a feature with a lot of benefits to the communities served by them. As maintainers design their resources in accordance with the research conducted, both representation and fit to potential uses can be improved. Such a local resource is most likely

useful. If it is too broad or not quite fitting, a smaller resource might be launched, catering to the needs that were previously unmet. While appealing in concept, this comes with a challenge. As the fit to a given community is improved, integration and sharing across contexts become more challenging. The underlying problem of too much and too messy data is only partly solved when too many, too different resources order the underlying data. The solution to that problem is integration based on standards. But here, again, the needs of many involved actors need to be considered. Throughout the interviews, a widely shared comic describing the problems of such standardization efforts was brought up, not just by one interviewee but by many.



Comic: Standards (XKCD)

Here, the usefulness to a few is in conflict with the potential of shared structures. To balance the distributed nature of biodata resources serving their communities, their integration seems to benefit from central coordination. Such coordination is yet another task on the plate of maintainers, and its scope can be quite extensive.

> *The number of slacks I'm on these days is kind of stupid. So there's a few big ones. So there's the Global Biodata Coalition, the GBC, and that is really trying to bring together the funders. So the people who fund these various resources to talk about their issues […]. There's the OBO Foundry, which is the Open Biomedical Ontologies Foundry, which is trying to bring together all the people who are coming up with ontologies to describe all of this different data, so that we don't, you know, as the XKCD cartoon has… […] So we keep trying to build connections to various networks in various places.*
>
> *(Sophie, Pos. 39-42)*

In this network of consortia, some have taken over the task of centrally coordinating certain aspects of resource maintenance and data aggregation. Directly addressing the problem described in the comic, the OBO Foundry tries to ensure that new ontologies only cover a space that is not yet filled. The value they provide to their members is structure and rules for the structured vocabularies that are used by resources. This collaborative work equips maintainers with a set of vocabularies they can implement to ensure that data can be integrated across resources.

> *So most of the ontologies we use have to be in the OBO Foundry, and they have to make an application to be in the OBO Foundry. And at that point, checks are done to make sure that ontology space is not covered already by another ontology, so we don't get duplications. And they will also make sure that the ontology, this provision… that is going to be maintained. It's not somebody's pet ontology. It's going to be useful in the biological ecosystem.*
>
> *(Hannah, Pos. 61)*

When coordinating across resources, stability is even more important than on the resource level, as large changes would most likely necessitate changes to the resources using the ontology. Additions of new terms might be a rather routine process, while larger, more structural changes require more extensive foresight and consideration. For maintainers contributing to these efforts, all levels need to be considered. Consequences for data aggregates, for data work and for the infrastructure facilitating the two all come with value conflicts that impact how changes can and should be implemented. Under coordination with subject experts, those changes can require extended deliberation.

> *We spent a long time arguing over… Or well, discussing what the word "allele" means, and what's the difference between an allele and a variant? And what does it mean to be a variant? And what's the precise agreed definition of this word? And can we use it consistently everywhere, you know? What does it mean to be a model of something?*
>
> *(Sophie, Pos. 23-24)*

Such ontological harmonization is crucial to resources describing model organisms, like the one Sophie is working on. As model organisms are studied with the aim of inferring their properties to others, often to humans, the ability to integrate data across these organisms is central. Ontologies allow such transfer, matching the description of one organism to the description of another. Just as resources constantly develop, so do the ontologies they use. The structure of data is reassessed to make sure that the description is not just correct but representative. However, in these considerations, the ways of valuing described in the preceding section are not omitted. Here, decisions are highly consequential and therefore require even more attention and, equally important, a nuanced understanding of the workings of resources, of affected research fields and of the work that any decision brings along. In addition, these decisions are not made in a vacuum, building on a long legacy that needs to be considered and, to some extent, helps to assure stability.

> *So the Gene Ontology has existed for 25 years, maybe 26. And so, obviously, we've got a lot of stuff in there that's legacy. […] We're also being much more rigorous about how we define things, how we structure things. So there are quite a lot of changes, and we've been doing a lot of changes around things where we've got redundancy in the ontology, where we represented a term in the molecular function ontology, and then we've represented it again in a slightly different way in the biological process ontology.*
>
> *(Hannah, Pos. 50)*

Beyond ontological integration, coordination can also enable the technical connection of resources. As terms are already shared, connecting the databases technically allows their users to move between them freely, finding and accessing even more data. While knowledge and even curation do not change in this case, new possibilities to connect and synthesize data are opened. Data become usable in a new way. When asked for possible improvements to the resources used, Liv highlights the value of such connections.

> *You get links to the databases, but you still have to find the sequences. […] You know, having a feature that would do that for you instead of having to go and manually find them all like that would be useful. Anything that streamlines those sorts of things is helpful. And of course, [the resource I use] doesn't have, for instance, links to like every tool that's available or every database.*
>
> *(Liv, Pos. 132-133)*

While programming capacities are often a scarce resource for maintainers, such integrations can provide increased usefulness. Depending on the technical features present, such an integration requires effort from both involved resources and a coordination of such efforts. The design of resources, which is often focused on openness, allows such work across teams and technical contexts.

> *We've had some people do pull requests into it – mostly when it's collaborations with other organizations to make improvements or improve an integration more.*
>
> *(Valentina, Pos. 84)*

By integrating in such a way, resources become reliant on other resources, making some of them more crucial to the sustainability of the field than others. Especially larger repositories, providing access to research papers and published data sets, take over some of the redundant work for smaller and more specialized resources. From a maintenance perspective, this means that workloads are reduced. However, the increased interconnectedness makes developing rules and infrastructures ever more complex. While this ideally results in consistency across systems, it also makes the long-term maintenance of those dependencies a crucial task for the stakeholders involved. As such, coordination also includes the assurance of funding and upkeep for those core resources.

> *We now have 52 of those data resources. I think there are more data resources in the world that could be core. But it's probably a number that tops out at something, like a hundred. 200 maybe. […] To put it negatively, if you took these core databases away, then many databases that are more specialists, that depend on them for their data input or for the tools or for some part of their process They would stop being viable.*
>
> *(Nico, Pos. 56)*

In relation to funding, but also in relation to the ontological and technological integration, coordination is a matter of consistently providing aggregates that serve both specialized research fields and biological research as a whole. Community, in this context, is not only provided by biodata

resources; these resources themselves are dependent on constant coordination and tight connections. While users value resources for their power to bridge distances within research fields, maintainers rely on large communities of peers to allow for large-scale integration and harmonization. It is therefore not surprising that such coordination seems to play a crucial role in their work. As a collective and as individuals, maintainers aim to ensure their work is valuable to as many as possible. In their collaboration data, data work and data infrastructures are valued through a generalized version of the primary registers of valuing. In their integrative efforts, they aim to combine locally stabilized valuation with ways of valuing that are shared across contexts. They make knowledge locally and globally valuable, useful and usable. This work is both determined by those interacting with their resources and by the maintainers' peers. To better understand this bridging of distances, the last section of this chapter will examine valuation and maintenance through a number of contexts in which maintainers operate.

6.3.3 Creating Infrastructure Contexts

Biodata resources, their maintainers, and those that interact with them are part of a number of contexts that are important to consider in this thesis. Contexts that themselves provide a number of advantages and shape, at times passively, at others actively, the work in and around data pipelines described so far. In this section, I do not aim to provide a conclusive list of such contexts, however, I bring together a few actors and groups involved in the creation of contexts. Starting small, I highlight the networks supporting many resources directly. Governance structures and research communities perceive and value resources from a short distance. They contribute and benefit, valuing accordingly. Beyond users and contributors, funders provide a monetary context, enabling resources as well as much of the research they aggregate. For many of them, value is tied to notions of impact and efficiency. Lastly, the local institutional context of maintainers and curators makes up the immediate surroundings. An often political space that can offer support or uncertainty, enabling maintainers or adding additional strain. As maintainers work through negotiations and stabilizations of value, these contexts add additional reference points, providing at times certainty and at times increased complexity. The actors shaping these contexts value the work of maintainers from their perspective, respectively, providing a sense of accountability. Starting locally:

> *The governance includes advisory structures and advisory boards. That meets, for the global core biodata resources, a requirement to have an advisory structure. But for the big database, they typically allow representative members of the community to be on these advisory structures.*
>
> *(Nico, Pos. 74-75)*

For many resources with dedicated funding, their actions are institutionally embedded in advisory structures like the ones mentioned by Nico. Made up of experts in the field or of peers, these boards represent the community served by the resource. Here, the way value is stabilized in the rules and infrastructures of a resource is put into conversation with the ways that its community values data and

data work. In less formalized settings, where resources lack funding and are not embedded in such formal structures, this task is still taken over by the community directly. As a whole, the research community represents the direct context in which resources and their maintainers operate. This context might be the most important, as the markers for value tied to these resources are closely connected to the communities' needs.

> *Everyone has to say, "This is my justification for running." You know, and this is one of the difficult things. Things don't always have to last. There'll be some databases that have a huge importance for a long period of time, but actually, the methods move on.*
>
> *(Nico, Pos. 78)*

Looking at the database landscape through Nico's statement creates a picture of constant change. Resources that were useful in the past might not be useful in the future. Yet, while demand lasts for a given research, its maintainers seem to strive more for reliability, structure and longevity. Balancing these outside pressures with measures of consistency and integration, Sophie cautions funders. She suggests staying with the resources currently supporting research, instead of funding what might be relevant in the future.

> *You know, it's hard because people want to fund the next big new thing. They don't usually like to fund infrastructure, and yet you need infrastructure to support the next big new thing. Like, you need it.*
>
> *(Sophie, Pos. 31)*

Whether matching the valuation of the community or not, funding infrastructures clearly make up a second important context. As resources aim to take over the data collection work for their users, they constantly struggle to secure the financial backing necessary, regardless of whether this comes in the form of paid hours as part of an unrelated employment, or as dedicated funds financing both small and large teams. Even for the digital infrastructures they rely on, maintainers are subject to a level of dependency.

> *So yes, our funders obviously have a very significant input into what we do. […] I mean, so there's two factors, right? […] They might say, "We're thinking we need to use this standard for whatever," so you need to deal with that. […] But more generally, we have to report to them. […] We can decide what we do and then justify it to them later, and then they can say: "Yes, that looks good." And so it's not like we're micromanaged, but there might be some specific things that they would like us to do.*
>
> *(Bejan, Pos. 17)*

As Bejan describes, funding comes with ideas of valuable resources, represented by specific features or broader directions. But value in the funding context is always also related to the funds provided. With funding bodies contributing on a larger scale across fields and sub-fields, questions of relevance and fit need to be answered differently.

> *The practical reason [for harmonization] is if we have to build a single backend. Because that's what the funders want to do. They don't want to pay for seven backends; they want to pay for one.*
>
> *(Sophie, Pos. 21)*

Usability in this context ideally means time saving for maintainers who are building tools across resources, and it means saving time for users who are able to connect and consider even more data, even more conveniently in their work. For funders supporting both the resources and the researchers, harmonization represents saving time twice, while supporting ever bigger, more academically impactful resources. As such, they value increasingly structured knowledge that is useful to as many as possible. Financing such a resource, however, also comes with a challenge to funders. While the openness and collaboration central to these resources afford them better tools for the research they support, funding open data infrastructure also means supporting the work of researchers who are funded by others. This effect is, of course, in line with more general ideals shared within the academic community, but it might be against the logic of impact-optimization under which some funders operate.

> *I don't think [the users] can imagine that there would be any reason why we wouldn't be funded. I don't. And we have been trying to instill into our community: it's not a given that we will always be here because the funding situation is dire at the moment. […] One of the things we are trying to do is think about ways to capture impact. And it's really hard because I would say that nearly all our impact can only be described qualitatively. It's really, really difficult to quantify.*
>
> *(Hannah, Pos. 70)*

Funders intending to fund only those resources that are impactful become yet another central element of the work of maintainers. To receive the support necessary, they constantly need to communicate the usefulness of something that seems to be a matter of principle for them. In their communication, they negotiate ideals of openness and control. As decentralized infrastructures, they aim to provide resources that support even unforeseen scientific progress by providing a wide array of data not strictly tied to current use cases. At the same time, they need to limit their scope, or at least communicate a limited scope, closer aligned with the research needs as identified by funders.

> *And so we spend a lot of time trying to ensure that we have continued funding in other ways, which is one really hard thing in biocuration. So that's a lot of sort of thankless work in some ways, despite all the pride I have. That is the funding aspect, which is a challenge and probably why we've seen a number of other resources go under in the 10 years that we've been around.*
>
> *(Valentina, Pos. 174)*

Dependency on funders is a vulnerability, especially for those resources that are both too big to be side projects and too small and specialized to be a core infrastructure. In their position, shifts in the academic and funding landscape can threaten the work maintainers and their teams are doing. For some, this means being prepared to move the resource beyond its current organizational structure in case the funding structures do not provide the needed support any longer. Some maintainers plan ahead for the case that funders challenge the usefulness of resources, to enable continued open access for their users.

> *A lot of our colleagues and friends at other institutions that have developed similar resources have had to go behind a licensed model. And we have been absolutely dedicated, we're never doing that to the point that any of us could spin up an instance of [the resource] on a personal laptop and make sure that this goes… like, as a group, we've really decided this will stay public forever.*
>
> *(Valentina, Pos. 172-173)*

As interviews took place at the beginning of 2025, around the second inauguration of Donald Trump as US president, accounts of longevity at times had a more explicit political element to them. As maintainers see funding and political support for their efforts under threat, they identify politics as another relevant context in which they operate[6]. Across a number of national contexts, they make out larger shifts in funding availability, making this more than just a US topic. Europe and North America remain the places with the most funding available, yet, through shifting priorities, funding in these geographies is declining – that is, in the perception of the interviewees. While much of the focus of this thesis is mostly on the local actors, whose valuation shapes resources, these national and international contexts need to be considered as well. As data infrastructures get re-evaluated in political discourses, the larger contexts shift. However, politics is also relevant beyond the national stage, as similar discourses might take place on the institutional level.

Universities and similar bodies provide the direct context in which maintainers and users alike accomplish their work. Here, too, political support for infrastructuring efforts can be a crucial factor.

> *You need to have buffering from political perturbation that threatens things. You need to have neutrality. So you don't want to be running in an institute that, if the institute changes its mind… It's not gonna run neuroscience projects anymore, it's gonna purely switch to behavioral studies that don't involve molecular neuroscience. If there's a really important database, it can't just suddenly change to be to being something completely different.*
>
> *(Nico, Pos. 82)*

Agency and independence within an institutional context are therefore important features, putting control in the hands of maintainers and their communities. But a hands-off approach can also mean

[6] Possible quotes on this matter are actively omitted, as interviewees seemed cautious to speak openly during the recordings.

that funding is not provided, and other work continues to be expected. Especially for small resources, institutional freedom means that the development and maintenance of biodata resources becomes a side project with limited pay, if any.

> *My institution didn't ask me to do this. I got no money from my institution to do this. It was very much a nights and weekends thing at first, and now they're like: "It's great." They like to house scholars who do things that become well-known.*
>
> *(Jan, Pos. 146)*

Depending on and operating within these larger contexts, maintainers try to provide their communities with relevant aggregated data to make knowledge accessible and to support further research. In coordination with their peers, they hope to construct a decentralized infrastructure that is sustainable and long-lasting, structuring biology in a consistent way. They design the data pipelines that select, aggregate and communicate data, while communicating to researchers, contributors and users alike to ensure they know how to best interact with the resource and to gather and synthesize their needs for continuous implementation. This puts them at the center of decisions that often include a variety of stakeholders, each with their own valuations. For those involved, this work is at times about community and personal benefits, involving non-epistemic rewards as much as any other. Together with those interacting with the resource and the data that flow through them, maintainers continuously work on biodata resources to ensure that knowledge is complete, correct and structured in a way that is trusted by all. By making sure that their databases are useful for day-to-day research, they allow that knowledge to serve researchers and practitioners, providing them with access to the knowledge and data they need. As designers of rules and builders of infrastructures, maintainers negotiate between the short-term and specialized needs of users, the wants of funders and the actions of peers, all while building a resource that achieves to be valuable in the present and in the future. They are always open to suggestions, contributions, collaboration and support, yet remain cautious to give up control completely. The data aggregates assembled, the data work accomplished, and the data infrastructures built are too valuable to them and ideally as valuable to the community they serve.

7. Discussion

This thesis explores biodata resources through the lens of valuation and practice. It focuses on specialized knowledge bases, the maintenance and data work connected to them. The valuation studies approach helps to better understand how and to what ends these digital infrastructures are developed, upheld and interacted with. As such, this thesis answers the call of Critical Data Studies scholars (among them Kitchin & Lauriault, 2018, p. 8) to closely examine the human and normative work behind data aggregation. Further extending historical scholarship on supportive infrastructures in the life sciences, this project focuses on the people working in and around biodata resources. This exploration builds on the notion of data work, be it closely tied to the data or placed at a structural level. Throughout the exploration of practices, these different kinds of work are systematized. Building on additional scholarship on maintenance and especially software maintenance, this thesis

extends notions of upkeep beyond technical aspects. Besides the upkeep of source codes, maintenance is described as a social practice that is crucial to the development and stabilization of value in human and technical infrastructures.

Building on valuation studies means centering registers of valuing as categories of independent evaluation and valorization. These registers are embedded in situated practice, and, once identified, provide a rich vocabulary for further analysis. By seeing practice as an enactment of value, processing, interaction, and infrastructuring become sites of negotiation and stabilization. Here, value is perceived and made, contested and combined. Across registers and at all times situated in contexts and relations, value is embedded in the data aggregates, the data work and digital infrastructures. To better grasp these entities, additional concepts from valuation studies are borrowed. When describing practices as embedded valuation constellations, positions, rules, and infrastructures become describable stabilizations of valuation. Emerging from the interviews, these constellations further enrich the enactment of value in practice.

The result is, on the one hand, a detailed overview of registers of valuing and, on the other, a systematized understanding of practices in and around biodata resources. In the analysis, the two are connected answering (RQ) how data are valued in and around resources, (SQ1) how these values are enacted in practice and (SQ2) how they are combined and negotiated. To accomplish this, the analysis moves from inside the data pipeline to the outside, where researchers, contributors and users interact with data and data work. Lastly, infrastructuring is explored as the outermost layer, in the form of maintenance, collaboration and context creation. Throughout the role of maintainers is highlighted, as they are key figures in the negotiation and stabilization of valuations. In this chapter, scholarly work, the theoretical concepts and empirical findings will be put into dialogue, starting with an overview of core findings, answering the research questions raised above. The chapter's focal point will be additional remarks on the positions involved in and central features of maintenance in the context of biodata resources. Here, the knowledgeable, tireless and caring work required in continuous development and upkeep will provide an alternative perspective on value-enactment and negotiation.

To make data movable beyond their context of origin, the work of preparation and description is necessary. This translational process is central to bridging the distances between data creation and use (Borgman & Groth, 2025). But bridging the distance can not just be seen with a focus on curation. Within the data pipelines explored, making data movable is similarly an exercise of finding data that might be valuable beyond their original purpose, preparing them in a way that this value is perceivable in a new context and communicating data and their value across time and space. This work is directly connected to the data; it happens with data. By selecting, curating and communicating data, they are made movable and perceivable. Closely tied to the original context of the data and any potential future use, this process aims to make understandable how data are created, what they are describing, and what their relations to reality might be. As data from different contexts are worked on in the same way, they are aggregated and integrated to form a larger body of knowledge serving more than just a single context. But throughout the interviews, these processes of data work are just

described in relation to the data, in relation to the context of origin and the potential contexts of use. Descriptions of what happens to and with data seem disembodied. In these accounts of data work, data are valued through the primary registers of value, especially in relation to their knowledge value and their usefulness elsewhere. The practices that are placed *in* the data pipeline are concerned with making data movable and moving it along – not much more. When simplified in such ways, this digital context through which data pass, the place of translation, is seen as guided mostly by curation rules and less by situational factors. This telling of the structured nature of curation, of work guided by ontologies and guidelines, has come up throughout the interviews, but not without critique.

While logical and clearly defined in theory, this work cannot be successfully achieved by machines. The output is highly structured and machine-readable, and yet humans are needed to make the data's context of creation computable. While humanly enacted, curation rules shape what should come out, and in which ways data should be described and prepared to make them and their value movable – academic value that is. What is striking in this process is that the rules, seemingly, take over the decision on how data is to be seen. While a human-driven process, the humans are mostly enacting the predefined steps in the pipeline.

The meeting of curation rules and human curators might lead to conflict. If the rules don't suffice or are not in line with the interests of those moving data along, then the messiness begins. The person moving the data is acting within their role. As a curator, they translate, describing and annotating in accordance with the guidelines, manuals and commonly shared terms. However, outside of the position, these rules and the prescribed practices might not always be the most valued approach. This is in line with Plantin's (2021) description of archival data workers who strive to find agency outside of the curatorial pipeline. When switching between positions, priorities shift, and for researchers or users who have limited time, not all the ways of valuing embedded in the position of "curator" might fit all the time. The interviewees shift between such positions again and again, describing the value of aggregated data and their struggles as contributors with potentially diverging goals.

While the inside of the data pipeline is structured by rules, the outside, the interaction, is structured by a limited number of positions that can be embodied. Each of the positions discussed in the empirical chapter comes with its own affordances. As researchers, publications can be made more fitting for the pipeline. As contributors, data can be moved along in very standardized ways and as users, predefined visualizations and formats make the data perceivable. So close to the data pipeline, the primary registers of valuing remain an important aspect. They are deeply entangled with the positions available. In the interaction with the data pipeline, however, there are not just rules and positions involved. Researchers, contributors and users are not neatly separated groups of people, and as such, their embodiments of the positions might not always be absolute.

For many of the resources discussed, researchers, curators, and users are overlapping groups. The result, it seems, is the emergence of additional ways to see data and data work as valuable. In the case of community curation, curators do not just curate for the sake of future use, but might additionally be motivated by non-epistemic academic rewards or engaged through community and personal

development. Some of these additional ways of valuing have become embedded in the positions. As humans interact with the data pipeline, data are valued in more than just the most important ways. They are made movable to enable science elsewhere, but also enable the bridging of social distance between members of a research community. What emerged from the interviews is not just positions tied to specific types of practices, but also how these positions are shifted and redefined to provide more than just engagement with aggregated data. Within the positions, registers of valuing are not just enacted but combined and, to some degree, negotiated. The degree of this negotiation, however, remains limited. That is because curators perform data work that is mostly structured by the rules within the pipeline. For example, interviewees talk about how certain fields need to be filled in, and how annotations that go beyond the scope of the curation manual might not fit. These infrastructural preconditions render the curated data *as valuable as possible* within the specified structure. Still, curation guidelines are not without room for movement, at times allowing workarounds or "additional comments." As such, rules are enacted in different ways, and predefined positions become embodied and tied to the individual ways of valuing.

The result is a constellation that values data and data work complexly. Aimed at making data movable, interaction with biodata resources seemingly enacts multiple ways of valuing. Curators might be able to choose what research should be in the collection, but are then guided by somewhat strict curatorial guidelines. Their engagement is driven both by personal benefits – academically, career-wise, personally – and by a sense of belonging and community. This happens in a stabilized manner, as more than data-focused ways of valuing are taken into consideration by maintainers as they work on resources. While designed to serve data-focused purposes for somewhat defined groups, the maintainers featured in this thesis continuously improve their resources with the communities they serve in mind. As such, they build not only digital data infrastructures but valuation infrastructures that provide a context for valuing of many kinds. Focussed, again, on the primary ways that data are perceived as valuable, these resources are built around the data pipeline. They are directed to enact data as complete, correct and structured, while doing so in a trustworthy way. Along the path of data from origin to use and re-use, these infrastructures are aimed at longevity. Their aim is to increase the usability of large amounts of research, sparing individual researchers from the need to aggregate and curate for themselves. However, as a place of curation and reconfiguration, their consistency is key. Biodata resources aim to re-value data in a reliable and predictable way, and they do so with a lasting impact. While the aim of this thesis was to look at the enactment of value, these infrastructures instead embed value in them, emerging as stabilizers of complex ways of valuing data as well as data work. In curation interfaces and collaboratively developed ontologies, in curation rules and prescribed positions, certain ways of valuing are put forward. Data are selected, made movable and communicated in predefined, infrastructured yet not value-free ways. Embedded in resources are ideas of *good* data and data work shared by certain fields that operate in close proximity to them. Embedded are the perceived or directly voiced needs and wishes of researchers, curators and users that interact with resources daily. And embedded are the priorities of maintainers, their peers and relevant stakeholders that shape and aim to shape the future of aggregated data.

While data work is often hidden, described as a service that transitions data from one context to another (Davies & Holmer, 2024), the infrastructures that enable and shape it are hidden just as much. While often described in the same vein, the work needed around the data pipeline presents a field of practices largely unconsidered. This ongoing maintenance not only enables the work within the pipeline, but also provides structure and rules, predefines positions and their responsibilities. Additionally, as part of maintenance, the digital infrastructures and their embedded values are constantly worked on. Here, rules are transformed into code and interfaces, deeply entangled with the entire path of the data and in constant recalibration of the potential ways of interaction. These infrastructures bridge between the positions and rules, while prioritizing certain ways of valuing along the way. But enabling movement between many contexts is not just a local task.

To make sure that the context where data are translated does not turn into many more, maintainers harmonize their efforts to make data move. The rules embedded and the positions they define are negotiated amongst collaborators concerned with similar goals. Through negotiation and even more translational work, a shared valuing of data is achieved. Here, valuing itself is made compatible, to allow for future smooth translations. Once embedded into the resources, this preemptive work enables data to be seen in many ways, making them movable for even farther distances – in some cases, even too far. As harmonization highlights similarities in valuing data, features that differ might be lost or at least deprioritized. The aggregate presents and enacts value that might in some cases go beyond what is there. This also needs to be considered. Whether collaborative or local, maintainers are the ones deciding how data enters, moves through and exits the aggregate. Throughout, they are themselves valued by others. Regardless of whether the resources evolve or stay the same, their value is constantly decided and enacted by their users and outside stakeholders. These groups are important, as use and funding make the maintenance worthwhile. As a proxy for the usefulness of resources, usage numbers are seen as one metric representing their value. Similarly, advisory boards and funding committees bring forward their own, yet sometimes opaque ways of valuing data aggregates, data work, and data infrastructures. For maintainers, communication with stakeholders therefore becomes a central practice. This provides maintainers a better understanding of how data and data work are valued by outsiders, and helps them adapt the infrastructure and their underlying rules and positions accordingly. However, ways of valuing can come in varied ways, as I have shown, so this task of sensing what might be important is a constant challenge. Balancing priorities, those in control collect and negotiate how others perceive their resource, the work that goes into it, and the data aggregated. Moving on a spectrum of openness and control, maintainers take into consideration short-term needs and long-term stability. This work includes not just the sensing, negotiation and implementation of these stakeholders' ways to value, but similarly important, the communication of the value of their own approach. As biodata resources value data in certain ways and structure data work accordingly, their shape and the positions and rules all need to be enacted as valuable themselves. The result is a complex entanglement of valuations:

At the center, data are evaluated and made valuable through data work. This data work is shaped by rules and infrastructure, which therefore enact some types of data work as more valuable than others. And lastly, the maintenance, description and presentation of infrastructures, once again, adds a third layer, valuing some infrastructure decisions over others. A complicated constellation, showing how value is enacted, combined and negotiated across and throughout a number of levels. From the valuing of data, to the valuing of data work and the valuing of data work infrastructures, all practices are embedded and shaped in relation. Making this matter even more complicated, these influences are mutual, multidirectionally connecting decisions about infrastructural changes with contributors' doubts about investing time in curating, and connecting larger shifts in research methods with the questioning of future funding. The result is a complex web of valuations deeply embedded in practices and the structures, positions and rules that stabilize them.

Over the last chapters, I have set out to show how data are valued in and around biodata resources, how they are valued differently across practices, and how valuations are combined and negotiated. In this chapter so far, I have laid the foundation to answer these questions. In doing so, I have at least scratched the surface of all the ways that the valuing by everyone involved influences the myriad of practices in and around biodata resources. Nothing more was my goal. Beyond, I want to answer these questions differently. Having given an account of the registers of valuing and their complex entanglement with practices of many kinds, my extended answer to the question of how data are valued is: knowledgeably, tirelessly, and with great care. Extending Davies & Holmer's notion of "the hidden work of biocuration" (2024), I will spend the rest of this chapter focusing on the relentless work of bio-maintenance.

First. Talking about software maintenance, Ojala & Cohn (2024) highlight the importance of knowledge to the sustainable upkeep and development of large-scale projects. As directions shift and long-term goals are revised, changes need to be integrated seamlessly to allow for future maintenance. This work of folding in new developments requires extensive knowledge of the project, its components and their functionality. I extend this perspective on knowledge beyond the code base. Maintenance for biodata resources is not just a matter of building digital tools and infrastructures, but also of integrating them with knowledge systems and larger academic contexts. This integration similarly requires extensive knowledge of the many moving parts explored above. As work in these resources is translational in nature, maintainers need to understand both the research community from which they are drawing data and the potential uses that their aggregated data could cater to. In addition, they build complex infrastructures, both digitally and epistemologically, that each require maintenance and constant enhancements. Integrated into larger networks of collaborators and embedded in academic and institutional contexts, maintainers need to consider both a complex landscape of values and the existing structures in place. Prioritizing between scarce resources and ongoing pressures, they decide between what can be changed easily and what needs to be changed urgently. Under the pressure to invent and evolve, they develop digital infrastructures that are ever new while maintaining consistency and ensuring future adaptability. Avoiding clutter, they

communicate openly, yet clearly, translating their deep understanding into relevant information for funders and users of all kinds. They are open for inquiry, yet cautious not to overburden others with additional tasks and communications. They are in control.

Second. The development and maintenance of biodata resources is ongoing and constant. As maintainers aim to fulfill data needs and feature wishes, their work does not stop. Navigating the various ways users, peers, and other stakeholders value their resources, their data work, and outputs makes their work a never-ending process of solving problems, both big and small. Like software (see Ensmenger, 2014), biodata resources are never "done." As infrastructures are constantly evolving, there is always a need, or at least a potential, to adapt. At any time, there are things to improve. In the interviews, as Hannah talks about the difference between "biologically correct" and "ontologically correct" (see p. 43) annotations, another step towards perfect representation is taken. Throughout the interviews, maintainers talk about the lengthy processes of deciding on descriptive terms and creating the most useful and correct ways to aggregate and communicate data. At the same time, their collections grow continuously, and new data are selected and integrated constantly. This is done to make the collection complete, to represent the latest state of knowledge and to ensure future relevance. Just like the fly stock center mentioned at the beginning of the state-of-the-art (see Bangham, 2019), biodata resources are under constant pressure to adapt and integrate, while ensuring high quality. But adaptation means re-developing, and re-developing always comes with consequences, with additional things to consider and with even more data work to be done. Similarly, integration requires coordination with peers, rebuilding of the ontologies used and changes to the structures of their systems. All of that happens under the consideration of potential additional workloads, the effects on users and researchers, and the funding rounds to come. Sophie's account of being in a "stupid" number of Slack channels reminds me of Linåker et al.'s (2024) problematization of maintainers' workload. While already focused on the data and maintenance work central to her position, communication with peers and stakeholders adds an additional toll. Like many maintainers of open-source projects, the maintainers interviewed here often take on additional responsibilities, sometimes without compensation and in most cases without secure funding. However, funding, next to usage, is crucial for the survival of these resources. In an analysis of shutdown repositories, Strecker et al. (2023) count that 64 out of 77 unmaintained resources ended their services due to institutional changes and funding cuts. An alarming rate, made even more alarming when considering that the number of resources staying afloat while struggling was more than six times higher than the number of those that recently closed, at least in the year 2000 (Merali & Giles, 2005). As translators between worlds, maintainers need to be aware of larger shifts. As shapers of data work, they balance at all times motivations and needs. As overseers of infrastructure, they attend to breakages, backlogs, and future research constantly. They combine what is known and anticipate what is to come. Their aim is to continually improve their systems, curation, and aggregates, while communicating with confidence the value of what is there. To communicate that usage, funding, and institutional backing are worthwhile, it seems their efforts can not stop.

Third. The maintainers interviewed for this thesis care. They listen to feature requests, consider them and decide the next steps. Their work is directed towards their users and the research they house, weighing constantly between what users want short-term and what is good for the aggregation long-term. They are not just interested in aggregating data, ensuring curation and building infrastructures; it seems they are trying to aggregate *good* data, ensure *good* curation and build *good* infrastructures. At this point, it should have become clear that this "good" is not an easy thing to achieve, because for the most part, "good" is different for everyone, every situation and every future imaginable. But like the definition of care, the central point of these efforts is to strive towards, and continue on, "as well as possible" (Fisher & Tronto, 1990, p. 40). Like data work can be conceptualized as many different things, so is the work of maintenance a process of caring for many parts at once. Not dependent on success but dependent on tireless and knowledgeable effort towards a better, sustainably relevant and long-lasting resource. Driven by what seems to be an intrinsic motivation towards their aggregative, integrative and communicative efforts, maintainers do more than data and maintenance work; they do data and maintenance care. In doing so, they maneuver between short-term and long-term, between mundane and innovative, considering all relevant factors in the decisions they make, finding the necessary balance between evolution and stability. In their decisions, the valuing of users, peers and stakeholders involved comes together, is brought into interaction, and stabilized through structural changes. This negotiative work is cognitive, requiring the consideration of many factors. But as decisions between ways of valuing are also decisions in relation to people, this work gains a social aspect. As maintainers engage with requests from users and in discussions with peers, as they argue their value to funders and communicate curation guidelines to curators, they need to consider more than just the practical implications of their work. Their work is emotional work. In many cases, the extensive engagement and work towards a better resource is a source of pride, providing a feeling of ownership. But with ownership, even if only emotional, comes the question of control and openness. While looking for support and for the integration of new voices into future developments, maintainers of the platform also need to ensure that those who get involved care as well. Making sure that contributions are either in good faith or managed through controlled routes, maintainers do not just work towards better data resources, but also towards what they deem "better" to be. Standing their ground, keeping everything going, and building on, maintainers keep their resources alive.

Knowledgeable, tirelessly, and with care, maintainers do what is central to making data move. As contexts of data creation and data use differ, translation is crucial to ensure movability. Similarly, as ideas of *good* data work differ between usage and curation, between publishing and funding, maintainers translate between ways of valuing that have just as much distance. Their practices are the enactments of important value judgements that close those conflicts, but they shape much more than just the aggregate. Influencing knowledge, its usefulness and usability, while a core priority, is not the only impact of maintenance. Building infrastructure, collaborating on rules, and assigning positions, maintainers' practices influence others' practices, and, as such, data valuation in and in interaction with the data pipeline. Their central roles put them in positions to shape the future data and

knowledge available, but it may also mean that the resources require their maintenance for a little longer.

> *Not everybody's thinking about how to make sure the project can move past its original creator/maintainer.*
>
> *(Louis, Pos. 58)*

As central figures who know every aspect of resources and their contexts, whose work never stops, despite outside pressure and whose engagement is driven by seemingly endless care for data and research at large, how can there be succession? If resources are run by individuals or ever-smaller groups that know and care, can a resource live on without them? If funding is limited to the minimum, who would take over this tireless work? Can maintainers ever stop?

8. Conclusion

Biodata resources are central to biology in the digital age. They provide aggregated data, bring together relevant research and summarize what is known in a field, about a certain entity or for targeted applications. As measurements have become increasingly digital, their potential to be used beyond their initial context of creation expands. By taking over the crucial work of data collection, curation and bundling, resources make research outputs easier to interact with, while simultaneously bridging epistemic distances. This travel is, however, not without prerequisites. Data are subjective. They are produced to represent a certain aspect of reality, and their descriptive value is in the eye of the beholder (Borgman, 2012, p. 1061). As data move, their beholders change and with them the way they are seen, interpreted and valued. Connecting the context of data creation and the context of data use, translational efforts like the curation discussed in this thesis are important. This data work helps bridge the distance, requiring constant re-evaluation by research producers or users. It is valuative work, making sure that data remain valuable along their path by preserving their connection to reality.

Data work, however, can be understood in multiple ways. Building on research around data work, data stewardship and digital labour (Baker & Yarmey, 2009; Irani, 2019; Karasti et al., 2010), different types of interaction with data need to be considered. As more than neutral practice, data are shaped by their curators, while curatorship itself is often shaped by working conditions, curatorial manuals, guidelines and hierarchical management (Plantin, 2018). In the context of biocuration, these different layers – the curation, the creation of rules and manuals, and managerial tasks – are often, though not always, discussed in one breath (e.g., by Leonelli, 2016). As resources are run on small budgets and by small teams, these different types of work flow into each other. They nonetheless serve different purposes. In exploring them separately, maintenance provides another lens ot talk about data work that engages with data only on the structural level (Borgman et al., 2016). As an ongoing, long-term, yet constantly evolving practice, maintenance work is crucial to the sustainability of biodata resources. In shifting research landscapes with ever-new methodical and content needs, resources need to stay up to date and structurally adapt, while not losing their focus on future data uses.

To better understand this multifaceted context and the many types of data work and maintenance practices, a valuation studies approach is used as the theoretical foundation of this thesis. Building on notions of care (Mol et al., 2010; Puig de la Bellacasa, 2011), this perspective allows for an exploration of different perspectives on data and data work. As situated, temporal and contextual, values are more than singular. When analyzing these subjective ways of evaluating and valorizing, registers of valuing describe shared categories of sense-making that move between people, contexts and situations (Heuts & Mol, 2013). Examining their enactment and negotiation in practice allows for a better understanding of what is done and why. However, the exploration goes beyond these broad categories and their impact in practice. As data work is spread out consistently over different parts of the resource, valuation is more than just situational. It is often, at least in part, stabilized. Curation manuals act as valuation rules; different people are prescribed different positions depending on their current tasks, and the resource itself acts as infrastructure to facilitate the movement of data (Waibel et al., 2021). These structures are not just structures for practice but structures of valuation, stabilizing, though not forcing, valuations of certain kinds. Building on the conceptual consideration of valuation constellations, this thesis has gained yet another aspect, not initially planned, but offered by the material itself.

As the central figures of the biodata resources described, maintainers made up the largest group of the sample for this project. Their knowledge enabled a deeper understanding of curating practices and, more importantly, their contexts. Focusing on field-specific resources that are community-curated allowed for the exploration of different positions that are jointly enacted by the interviewees. This was especially the case for community curators, who move constantly between being research providers, contributors and users of the resource. They make up the second interviewee group. Similarly, maintainers who contribute to larger consortia to facilitate field-wide communication move between positions and form the third group interviewed for this thesis. The nine interviews provide the basis for the analysis and shape the results accordingly.

Extending the theoretical valuation studies concepts, the analysis firstly focused on identifying shared registers of valuing that shape how maintainers and users make sense of data and data work. These registers range from closely connected to data to more personal and communal. Primary registers identified are concerned with, for one, the value of knowledge, its completeness, correctness and structure as well as the trust in the aggregate. Just as data are interpreted differently throughout contexts, so are these descriptors of data. The category might be shared, but the understanding of the values behind it might differ. Further primary registers of valuing are concerned with the usefulness for research and downstream use, as well as with the usability and ease of use offered by the aggregate as such and the underlying interface. All of these registers are closely linked to the data, and often embedded in general evaluations of biodata resources. However, these ways of valuing are not the only ones. Secondary registers, in contrast, focus on the personal and institutional benefits that researchers and users may gain from their interaction. Non-epistemic rewards and personal development are named here next to perceptions of community building and collaboration. Lastly,

ideological questions of openness and control run parallel as yet another register of valuing, shaping a number of decisions throughout.

The main contribution of this project is the systematization of practices across various types and the entanglement of valuation within them. The connection of value and practice across positions offers a detailed perspective on the complex interactions that shape data and data work. By differentiating between practices in the data pipelines, in interaction with the pipeline, and shaping the pipeline, this thesis provides a distinction of positions that makes visible how priorities are distributed and how subsequent decisions shape data movement. Within the pipeline, when data are selected, curated and communicated, the primary registers of value remain salient. What is done here is done in accordance with elaborate rules or through predefined exceptions. Here, the data are valued.

In publishing research, when contributing time to curation, and when using the resources, practices are shaped differently. Those interacting have more agency, yet their actions are still shaped by their positions. As researchers, curators or users, their access to data is constrained differently each time, even as they switch between functions. As human enactors of these positions, their ways of making sense of data and data work go beyond the primary registers of valuing. Here, curation is seen, for example, in relation to academic impact or feelings of pride. Similarly, research publication and usage are not just shaped by the value of the data themselves. As positions evolve, these multiple ways of valuing become a stabilized part of them. By design, interaction is about more than data.

Moving to maintenance, these varied ways of valuing become more explicit. As builders of digital infrastructures that facilitate valuing, maintainers consider more than just the movement of data throughout. Considering user needs and curator capacities, they shape digital tools, curation manuals, and data strategies. Their work is aimed at making data movable, but also to facilitate data work more generally. Practices of infrastructuring at this level happen both locally and in collaboration with peers. Ontologies are developed to allow the movement of data even across contexts with large distances. A translational effort, enabled through the coordination between resources or in consortia. This work shapes epistemological and technical integration, but is, by definition, a bridging act between contexts. Maintenance is at the heart of value negotiation, considering user groups, curation contexts, and peer exchange. Maintainers make sure that their resources provide valuable data and valuable data work. Their work is a constant process of sensing, implementing and communicating ways of valuing. Communicating to users is central to making sure their efforts are worthwhile, while communicating to funders and other stakeholders is central to keeping the resource and the needed data work going.

Under the pressure to upkeep resources that provide useful and usable data, and as mediators between distant needs, maintainers do more than just ongoing data work. Their negotiation and stabilization of value is emotional work as well. Extending work on software maintenance and maintainer fatigue (Ensmenger, 2014; Linåker et al., 2024), the work can be described as entangled in knowledge, as tireless and as a continuous caring practice. As, at times, sole overseers of the digital space they create, and as connectors between academic contexts, their knowledge is a key factor in making data

and data work valuable. With maintenance as an ongoing practice and with software as a never-finished building material, their work is continuous, too. It is not just about keeping what is there running, but as much about adapting themselves, their resources, and the underlying data structures to current and potential future needs. Always striving towards the best possible representation, the most useful data and the most usable resource, their engagement is caring. Striving towards, deciding between, and moving beyond to make biodata resources that last valuably.

Their central roles put maintainers under pressure. While their position are often under threat, or officially non-existent, their work is impactful. The combination of high responsibility and insecure pay makes their limited visibility additionally straining. Especially in the cases of the focused resources discussed throughout this thesis, this combination seems to be a widely shared experience. Here, especially, maintenance is often only one of the practices that people engage in, as positions and responsibilities fall together. The importance of these central figures raises further questions that extend beyond the scope of this thesis. As maintainers are underrecognized for their integral contributions to often widely used resources, the sustainability of their care needs to be seen as a matter of dedication. Between insecure funding and high responsibility, these positions seem tied to those who care to fill them, but succession and future relief are in question. Here, further research is needed regarding perceptions of sustainability among different stakeholders and on the longevity of resources that manage to distribute responsibility across larger groups.

Closer tied to the scope of the thesis, looking at practices in a systemic way is not just useful for knowledge bases in biological research. This distribution of positions and responsibilities, the valuing across a wide array of practices and the infrastructural stabilization of value judgment are applicable elsewhere as well. As described by Linåker et al. (2024), more-than-cognitive work is a shared experience with other open source maintainers, and the complexity of different types of data work is not a solely life science phenomenon (Borgman et al., 2016). Examining research data infrastructure more broadly, the differentiation of data, data work, and data work infrastructures can be a valuable tool for research on data valuation. Here, valuation constellations may also be a fruitful lens for better understanding how valuation is stabilized and enacted across different types of positions and practices. The systematization of registers of valuing can further enhance the exploration of data work at large – especially in consideration of primary data-focused valuation registers and secondary registers, which bring into focus the more-than-data-driven factors. As such, this thesis contributes not just to the exploration of data work and maintenance in biology but also to larger discourses on the aggregation and value-driven usage of data through digital infrastructures. Following the call of critical data scholars, the case study of biodata resources makes visible what is usually hidden behind the user interfaces of online platforms. Due to the inherently open nature of these research infrastructures, the work behind data aggregation becomes visible, and its structures emerge. More importantly, this thesis shows the value-laden nature of these practices, as well as the constellations they are embedded in. Not just is the value of data in the eye of the beholder, but data pipelines and the positions, rules and infrastructures that make them a reality are all the result of varied valuation.

My contribution to this discourse is to lay out how diverse these valuations can be. Their complex entanglements, contradictions and active stabilizations just add another layer.

In answering the research questions, I have provided a better understanding of the important role that valuing plays in practices of data work and in the rules, positions and infrastructures that shape them. The entanglement of data in more than just their context of origin and their potential contexts of use has to be considered, highlighting additionally a third context – their context of aggregation, integration, and communication. Shaped by maintainers, the practices in this third context shape data, but similarly influence people and research communities at large. Taken up with a sense of responsibility, this work is care work; it is tireless work and is built on deep knowledge, moving data across distances of so many kinds.

Bibliography

Asdal, Kristin, Doganova, Liliana, & Fochler, Maximilian. (2024). Valuation Studies as a Frame in Sts. In U. Felt & A. Irwin (Eds.), *Elgar Encyclopedia of Science and Technology Studies* (pp. 79-86). Edward Elgar.

Baker, Karen S., & Yarmey, Lynn. (2009). Data Stewardship: Environmental Data Curation and a Web-of-Repositories. *International Journal of Digital Curation*, *4*(2), 12-27. https://doi.org/10.2218/ijdc.v4i2.90

Ball, Stephen J. (1994). Researching inside the State: Issues in the Interpretation of Elite Interviews. In *Researching Education Policy* (pp. 113-125). Routledge.

Bangham, Jenny. (2019). Living Collections: Care and Curation at Drosophila Stock Centres. *BJHS Themes*, *4*, 123-147. https://doi.org/10.1017/bjt.2019.14

Borgman, Christine L. (2012). The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059-1078. https://doi.org/10.1002/asi.22634

Borgman, Christine L., & Bourne, Philip E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.42eec111

Borgman, Christine L., Darch, Peter T., Sands, Ashley E., & Golshan, Milena S. (2016). The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy. https://doi.org/10.48550/ARXIV.1611.00055

Borgman, Christine L., & Groth, Paul. (2025). From Data Creator to Data Reuser: Distance Matters. *Harvard Data Science Review*, *7*(2). https://doi.org/10.1162/99608f92.35d32cfc

Bowker, Geoffrey C., & Star, Susan Leigh. (1999). *Sorting Things Out: Classification and Its Consequences* (1 ed.). MIT Press. https://doi.org/10.7551/mitpress/6352.001.0001

boyd, danah, & Crawford, Kate. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662-679. https://doi.org/10.1080/1369118X.2012.678878

Burge, S., Attwood, T. K., Bateman, A., Berardini, T. Z., Cherry, M., O'Donovan, C., Xenarios, L., & Gaudet, P. (2012). Biocurators and Biocuration: Surveying the 21st Century Challenges. *Database (Oxford)*, *2012*, bar059. https://doi.org/10.1093/database/bar059

Cardoso Llach, Daniel. (2019). Tracing Design Ecologies: Collecting and Visualizing Ephemeral Data as a Method in Design and Technology Studies. In *Digitalsts* (pp. 451-471). Princeton University Press. https://www.degruyterbrill.com/document/doi/10.1515/9780691190600-031/html

Chen, C., Yaari, Z., Apfelbaum, E., Grodzinski, P., Shamay, Y., & Heller, D. A. (2022). Merging Data Curation and Machine Learning to Improve Nanomedicines. *Adv Drug Deliv Rev*, *183*, 114172. https://doi.org/10.1016/j.addr.2022.114172

Chen, Q., Britto, R., Erill, I., Jeffery, C. J., Liberzon, A., Magrane, M., Onami, J. I., Robinson-Rechavi, M., Sponarova, J., Zobel, J., & Verspoor, K. (2020). Quality Matters: Biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases. *Genomics Proteomics Bioinformatics*, *18*(2), 91-103. https://doi.org/10.1016/j.gpb.2018.11.006

D'Ignazio, Catherine, & Klein, Lauren F. (2020). Seven Intersectional Feminist Principles for Equitable and Actionable Covid-19 Data. *Big Data & Society*, *7*(2), 2053951720942544. https://doi.org/10.1177/2053951720942544

Dalton, Craig, & Thatcher, Jim. (2014). What Does a Critical Data Studies Look Like, and Why Do We Care? *Society and Space*. https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Davies, Sarah R. (2025). Working in Biocuration: Contemporary Experiences and Perspectives. *Database (Oxford)*, *2025*. https://doi.org/10.1093/database/baaf003

Davies, Sarah R., & Holmer, Constantin. (2024). Care, Collaboration, and Service in Academic Data Work: Biocuration as 'Academia Otherwise'. *Information, Communication & Society*, *27*(4), 683-701. https://doi.org/10.1080/1369118x.2024.2315285

Denis, Jérôme, & Pontille, David. (2017). Beyond Breakdown: Exploring Regimes of Maintenance. *Continent*(6.1), 13-17. https://hal.archives-ouvertes.fr/hal-01509617

Doganova, Liliana, Giraudeau, Martin, Helgesson, Claes-Fredrik, Kjellberg, Hans, Lee, Francis, Mallard, Alexandre, Mennicken, Andrea, Muniesa, Fabian, Sjögren, Ebba, & Zuiderent-Jerak, Teun. (2014). Valuation Studies and the Critique of Valuation. *Valuation Studies*, *2*(2), 87-96. https://doi.org/10.3384/vs.2001-5992.142287

Drysdale, R., Cook, C. E., Petryszak, R., Baillie-Gerritsen, V., Barlow, M., Gasteiger, E., Gruhl, F., Haas, J., Lanfear, J., Lopez, R., Redaschi, N., Stockinger, H., Teixeira, D., Venkatesan, A., Elixir Core Data Resource, Forum, Blomberg, N., Durinx, C., & McEntyre, J. (2020). The Elixir Core Data Resources: Fundamental Infrastructure for the Life Sciences. *Bioinformatics*, *36*(8), 2636-2642. https://doi.org/10.1093/bioinformatics/btz959

Dussauge, Isabelle, Helgesson, Claes-Fredrik, Lee, Francis, & Woolgar, Steve. (2015). On the Omnipresence, Diversity, and Elusiveness of Values in the Life Sciences and Medicine. In I. Dussauge, C.-F. Helgesson, & F. Lee (Eds.), *Value Practices in the Life Sciences and Medicine* (pp. 1-28). Oxford University Press. https://academic.oup.com/book/10636/chapter/158640108

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science Friction: Data, Metadata, and Collaboration. *Soc Stud Sci*, *41*(5), 667-690. https://doi.org/10.1177/0306312711413314

Ensmenger, Nathan. (2014). When Good Software Goes Bad: The Surprising Durability of an Ephemeral Technology. MICE (mistakes, ignorance, contingency, and error) conference,

Fisher, Berenice, & Tronto, Joan. (1990). Toward a Feminist Theory of Caring. *Circles of care: Work and identity in women's lives*, *7*(5), 35-92.

Forseth, Ulla, Clegg, Stewart, & Røyrvik, Emil André. (2019). Reactivity and Resistance to Evaluation Devices. *Valuation Studies*, *6*(1), 31-61. https://doi.org/10.3384/vs.2001-5992.196131

Galletta, Anne. (2013). *Mastering the Semi-Structured Interview and Beyond*. New York University Press. https://doi.org/doi:10.18574/nyu/9780814732939.001.0001

Gillespie, Tarleton. (2010). The Politics of 'Platforms'. *New Media & Society*, *12*(3), 347-364. https://doi.org/10.1177/1461444809342738

Graham, Stephen, & Thrift, Nigel. (2007). Out of Order:Understanding Repair and Maintenance. *Theory, Culture & Society*, *24*(3), 1-25. https://doi.org/10.1177/0263276407075954

Greeson, Emma, Laser, Stefan, & Pyyhtinen, Olli. (2020). Dis/Assembling Value: Lessons from Waste Valuation Practices. *Valuation Studies*, *7*(2), 151-166. https://doi.org/10.3384/vs.2001-5992.2020.7.2.151-166

Gregg, Melissa, & Andrijasevic, Rutvica. (2019). Virtually Absent: The Gendered Histories and Economies of Digital Labour. *Feminist Review*, *123*(1), 1-7. https://doi.org/10.1177/0141778919878929

Helgesson, Claes-Fredrik, & Muniesa, Fabian. (2013). For What It's Worth: An Introduction to Valuation Studies. *Valuation Studies*, *1*(1), 1-10. https://doi.org/10.3384/vs.2001-5992.13111

Heuts, Frank, & Mol, Annemarie. (2013). What Is a Good Tomato? A Case of Valuing in Practice. *Valuation Studies*, *1*(2), 125-146. https://doi.org/10.3384/vs.2001-5992.1312125

Huang, Hong. (2015). Domain Knowledge and Data Quality Perceptions in Genome Curation Work. *Journal of Documentation*, *71*(1), 116-142. https://doi.org/10.1108/JD-08-2013-0104

International Society for Biocuration. (2018). Biocuration: Distilling Data into Knowledge. *PLoS Biol*, *16*(4), e2002846. https://doi.org/10.1371/journal.pbio.2002846

Irani, Lilly. (2019). Justice for Data Janitors. In S. Marcus & C. Zaloom (Eds.), *Think in Public: A Public Books Reader* (pp. 23-40). Columbia University Press. https://www.degruyterbrill.com/document/doi/10.7312/marc19008-003/html

Karasti, Helena, Baker, Karen S., & Millerand, Florence. (2010). Infrastructure Time: Long-Term Matters in Collaborative Development. *Computer Supported Cooperative Work (CSCW)*, *19*(3), 377-415. https://doi.org/10.1007/s10606-010-9113-z

Kitchin, Rob. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures &Amp;Amp; Their Consequences*. SAGE Publications Ltd. https://methods.sagepub.com/book/mono/the-data-revolution/toc

Kitchin, Rob, & Lauriault, Tracey P. (2015). Small Data in the Era of Big Data. *GeoJournal*, *80*(4), 463-475. http://www.jstor.org/stable/44076310

Kitchin, Rob, & Lauriault, Tracey P. (2018). Toward Critical Data Studies Charting and Unpacking Data Assemblages and Their Work. In *Thinking Big Data in Geography : New Regimes, New Research*. University of Nebraska Press. https://research.ebsco.com/linkprocessor/plink?id=7596ac7f-bc4f-3a82-a1b1-be7b46ed0838

Law, John. (2004). *After Method : Mess in Social Science Research*. Routledge, Taylor & Francis Group. https://ubdata.univie.ac.at/AC04577504

Lehman, M. M., Ramil, J. F., Wernick, P. D., Perry, D. E., & Turski, W. M. (1997, 5-7 Nov. 1997). Metrics and Laws of Software Evolution-the Nineties View. Proceedings Fourth International Software Metrics Symposium,

Leonelli, Sabina. (2009). Centralising Labels to Distribute Data: The Regulatory Role of Genomic Consortia. In *The Handbook of Genetics & Society* (pp. 495-511). Routledge.

Leonelli, Sabina. (2014a). Data Interpretation in the Digital Age. *Perspect Sci*, *22*(3), 397-417. https://doi.org/10.1162/POSC_a_00140

Leonelli, Sabina. (2014b). What Difference Does Quantity Make? On the Epistemology of Big Data in Biology. *Big Data Soc*, *1*(1). https://doi.org/10.1177/2053951714534395

Leonelli, Sabina. (2016). *Data-Centric Biology : A Philosophical Study*. The University of Chicago Press. https://ubdata.univie.ac.at/AC14549326

Leonelli, Sabina, & Ankeny, Rachel A. (2012). Re-Thinking Organisms: The Impact of Databases on Model Organism Biology. *Stud Hist Philos Biol Biomed Sci*, *43*(1), 29-36. https://doi.org/10.1016/j.shpsc.2011.10.003

Leonelli, Sabina, Diehl, Alexander D., Christie, Karen R., Harris, Midori A., & Lomax, Jane. (2011). How the Gene Ontology Evolves. *BMC Bioinformatics*, *12*(1), 325. https://doi.org/10.1186/1471-2105-12-325

Levin, Nadine, & Leonelli, Sabina. (2017). How Does One "Open" Science? Questions of Value in Biological Research. *Sci Technol Human Values*, *42*(2), 280-305. https://doi.org/10.1177/0162243916672071

Linåker, Johan, Link, Georg, & Lumbard, Kevin. (2024). Sustaining Maintenance Labor for Healthy Open Source Software Projects through Human Infrastructure: A Maintainer Perspective. Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement,

Martin, Aryn, Myers, Natasha, & Viseu, Ana. (2015). The Politics of Care in Technoscience. *Social Studies of Science*, *45*(5), 625-641. https://doi.org/10.1177/0306312715602073

Meier, Frank, Peetz, Thorsten, & Waibel, Désirée. (2017). Bewertungskonstellationen. Theoretische Überlegungen Zur Soziologie Der Bewertung. *Berliner Journal für Soziologie*, *26*(3-4), 307-328. https://doi.org/10.1007/s11609-017-0325-7

Merali, Zeeya, & Giles, Jim. (2005). Databases in Peril. *Nature*, *435*(7045), 1010-1011. https://doi.org/10.1038/4351010a

Mol, Annemarie. (2008). *The Logic of Care: Health and the Problem of Patient Choice*. Routledge.

Mol, Annemarie, Moser, Ingunn, & Pols, Jeannette. (2010). Care: Putting Practice into Theory. In M. Annemarie, M. Ingunn, & P. Jeannette (Eds.), *Care in Practice* (pp. 7-26). transcript Verlag. https://doi.org/doi:10.1515/transcript.9783839414477.7

Müller, Heiko, & Naumann, Felix. (2003). *Data Quality in Genome Databases*.

Murillo, Luis Felipe R. (2023). How to Avoid the "Infrastructural Blues"? Studying-While-Caring for Data Stewardship. *Annals of Anthropological Practice*, *48*(1), 36-51. https://doi.org/10.1111/napa.12208

Nadim, Tahani. (2016). Data Labours: How the Sequence Databases Genbank and Embl-Bank Make Data. *Science as Culture*, *25*(4), 496-519. https://doi.org/10.1080/09505431.2016.1189894

Ojala, Mace, & Leavitt Cohn, Marisa. (2024). Software Maintenance as Materialization of Common Knowledge. *Engaging Science, Technology, and Society*, *9*(3). https://doi.org/10.17351/ests2023.1325

Pinel, Clémence, Prainsack, Barbara, & McKevitt, Christopher. (2020). Caring for Data: Value Creation in a Data-Intensive Research Laboratory. *Social Studies of Science*, *50*(2), 175-197. https://doi.org/10.1177/0306312720906567

Plantin, Jean-Christophe. (2018). Data Cleaners for Pristine Datasets: Visibility and Invisibility of Data Processors in Social Science. *Science, Technology, & Human Values*, *44*(1), 52-73. https://doi.org/10.1177/0162243918781268

Plantin, Jean-Christophe. (2021). The Data Archive as Factory: Alienation and Resistance of Data Processors. *Big Data & Society*, *8*(1). https://doi.org/10.1177/20539517211007510

Puig de la Bellacasa, María. (2011). Matters of Care in Technoscience: Assembling Neglected Things. *Soc Stud Sci*, *41*(1), 85-106. https://doi.org/10.1177/0306312710380301

Quaglia, Federica, Balakrishnan, Rama, Bello, Susan M., & Vasilevsky, Nicole. (2022). Conference Report: Biocuration 2021 Virtual Conference. *Database*, *2022*. https://doi.org/10.1093/database/baac027

Ribeiro, Barbara, Meckin, Robert, Balmer, Andrew, & Shapira, Philip. (2023). The Digitalisation Paradox of Everyday Scientific Labour: How Mundane Knowledge Work Is Amplified and Diversified in the Biosciences. *Research Policy*, *52*(1), 104607. https://doi.org/10.1016/j.respol.2022.104607

Rodriguez-Esteban, R. (2015). Biocuration with Insufficient Resources and Fixed Timelines. *Database (Oxford)*, *2015*. https://doi.org/10.1093/database/bav116

Rubin, Daniel L, Lewis, Suzanna E, Mungall, Chris J, Misra, Sima, Westerfield, Monte, Ashburner, Michael, Sim, Ida, Chute, Christopher G, Storey, Margaret-Anne, & Smith, Barry. (2006). National Center for Biomedical Ontology: Advancing Biomedicine through Structured Organization of Scientific Knowledge. *Omics: a journal of integrative biology*, *10*(2), 185-198.

Russell, Andrew L., & Vinsel, Lee. (2016). Innovation Is Overvalued. Maintenance Often Matters More | Aeon Essays. *Aeon*. https://aeon.co/essays/innovation-is-overvalued-maintenance-often-matters-more

Russell, Andrew L., & Vinsel, Lee. (2018). After Innovation, Turn to Maintenance. *Technology and Culture*, *59*(1), 1-25. https://www.jstor.org/stable/26803881

Star, Susan Leigh. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377-391. https://doi.org/10.1177/00027649921955326

Strecker, Dorothea, Pampel, Heinz, Schabinger, Rouven, & Weisweiler, Nina Leonie. (2023). Disappearing Repositories: Taking an Infrastructure Perspective on the Long-Term Availability of Research Data. *Quantitative Science Studies*, *4*(4), 839-856. https://doi.org/10.1162/qss_a_00277

Tempini, Niccolò. (2017). Till Data Do Us Part: Understanding Data-Based Value Creation in Data-Intensive Infrastructures. *Information and Organization*, *27*(4), 191-210. https://doi.org/10.1016/j.infoandorg.2017.08.001

Tempini, Niccolò. (2021). Data Curation-Research: Practices of Data Standardization and Exploration in a Precision Medicine Database. *New Genetics and Society*, *40*(1), 73-94. https://doi.org/10.1080/14636778.2020.1853513

Vatin, François. (2013). Valuation as Evaluating and Valorizing. *Valuation Studies*, *1*(1), 31-50. https://doi.org/10.3384/vs.2001-5992.131131

Vila-Henninger, Luis, Dupuy, Claire, Van Ingelgom, Virginie, Caprioli, Mauro, Teuber, Ferdinand, Pennetreau, Damien, Bussi, Margherita, & Le Gall, Cal. (2024). Abductive Coding: Theory Building and Qualitative (Re)Analysis. *Sociological Methods & Research*, *53*(2), 968-1001. https://doi.org/10.1177/00491241211067508

Wagenknecht, Susann, Kocksch, Laura, Laser, Stefan, & Kühnen, Ann-Kristin. (2024). Valuing Data: Attaching Online Data to Stakes, Selves, and Other Data. *Valuation Studies*, *11*(1), 60-90. https://doi.org/10.3384/vs.2001-5992.2024.11.1.60-90

Waibel, Désirée, Peetz, Thorsten, & Meier, Frank. (2021). Valuation Constellations. *Valuation Studies*, *8*(1), 33-66. https://doi.org/10.3384/VS.2001-5992.2021.8.1.33-66

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand Jan, Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas, Boiten, Jan-Willem, da Silva Santos, Luiz Bonino, Bourne, Philip E., Bouwman, Jildau, Brookes, Anthony J., Clark, Tim, Crosas, Mercè, Dillo, Ingrid, Dumon, Olivier, Edmunds, Scott, Evelo, Chris T., Finkers, Richard, . . . Mons, Barend. (2016). The Fair Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Zakharova, Irina, & Jarke, Juliane. (2024). Care-Ful Data Studies: Or, What Do We See, When We Look at Datafied Societies through the Lens of Care? *Information, Communication & Society*, *27*(4), 651-664. https://doi.org/10.1080/1369118x.2024.2316758

Zuboff, Shoshana. (2019). *The Age of Surveillance Capitalism : The Fight for the Future at the New Frontier of Power*. Profile Books.

Appendix 1: Interview guidelines

Exploratory Interviews	**Maintainers**	**User-curators**	**Consortia Members**
Intro			
[Introduce yourself and the research project without giving too much information.]	[Greetings. Ask for consent to record. Start recording once consent is given.] [Inform the interviewee about recording use, storage, publication, and possibilities to strike parts or the entire interview – both right after the interview or until publication. Start the interview once consent is given again.] [Introduce yourself and the project briefly, without giving too much information.]		
Phase 1: Types of curation & personal position	***Phase 1: Personal work and path***		
If you are looking at the resources for biodata, how would you describe the landscape? *What categories would you use?* *Are there resources that you perceive as more or less trustworthy?* *What kinds of people/roles work on resources?*	What does your work on the resource look like? *What are your day-to-day tasks?* *How important is it that others can benefit from your work?* *What is your motivation?*	What does your contribution to resources look like? *What does curation entail?* *How did you start contributing?* *What is your motivation to contribute?*	What does your work in consortia look like? *How did you enter this space, and why are resources so important?* *What does your engagement entail?*
Phase 2: Curation quality	***Phase 2: Curation practices and kinds of curation***		
What, for you, is a sign of quality in a curated resource? *Do those contributing/ using the data have different markers for quality?* *How do different processes of curation influence quality?*	How is curation structured by your resource? *What factors influence change in curation processes?* *Who is this data/output useful for?* What is the output of your work?	How do the resource and its data tie into your usual work? *How does the data change by being in the resource?* *How much do you work with the contributions of others?* What do you get out of the resource?	How does curation contribute to the work of resources? *How do you see different approaches to curation?* What drives change in the development and maintenance of resources? What output do resources provide?
Phase 3: Personal work	***Phase 2: Valuing biodata resources*** [Open section by giving a brief overview of the topic of the thesis, while highlighting the more than economic nature of value.]		
Can you talk a little about your work on resources? *What are you (and the team) looking at/ prioritizing?* *Who are the stakeholders that drive changes in the development?* How important is it for a resource to communicate curation?	How are resources valuable? *In what ways is this data important/valuable?* *Are data more valuable by being in the resource?* *How important are funding and impact?* How do you think different user groups value the resource?	How are resources valuable? *In what ways is this data important/valuable?* *Are data more valuable by being in the resource?* Do you think your perspective on the resources & data is different from the perspective of maintainers?	How are resources valuable? *In what ways is this data important/valuable?* *Are data more valuable by being in the resource?* *How important are funding and impact?* What future benefits and trends do you imagine?
Outro:			
Was there anything you were expecting to speak about? Do you have any questions for me?			

Appendix 2: Reflection on AI use

AI Tool	Work Phase	Usage	Comments and Reflections
AI Transcription Software	Transcription of interviews	Initial transcription and time-stamping were performed by two AI transcription services with good privacy terms	Throughout the transcription process, the software was swapped out, as the quality of the first software was far inferior, requiring extensive editing.
U:AI	Transcription of interviews	In cases where the automatic transcription failed and biological terms were not identifiable by me, these could be identified by providing the LLM with the surrounding text context, as well as a phonetic spelling of the word in question.	This approach proved much more fruitful than a Google search, as the LLM was able to use the context to identify words that not only sound similar but also fit the logical context. All words identified through this process still had to be checked through a Google search.
	Interviwee search	In collaboration with the LLM, a search strategy was developed to identify relevant resources and interviewees. The conversation resulted in a list of search terms and keywords.	The developed search strategy was somewhat effective, although repositories that collect lists of different biodata resources ultimately played a larger role in sample building.
Grammarly	Editing	Grammarly Pro and its desktop integration supported the editing by highlighting spelling and punctuation mistakes.	This tool, while effective for spelling and punctuation correction, also suggests rather invasive stylistic changes. These were not used.

This table is structured according to the AI regulations of the STS Department at the University of Vienna. Their blueprint table is based on the "Documentation Table" developed by the Writing Center of UniGraz (available here).