

Documenting Digital Legacy:
Provenance, Paradata, and the Practices of Three
National Libraries

FRANCISCO FERRER RUIZ
MANUELA MERCADO CAMPERO



UNIVERSITY
OF BORÅS

© **Francisco Ferrer Ruiz & Manuela Mercado Campero**
Partial or full copying and distribution of the material in this thesis
without permission is forbidden.

English title: Documenting the Digital Legacy: Provenance, Paradata, and the Practice of Three National Libraries

Author(s): Francisco Ferrer Ruiz & Manuela Mercado Campero

Completed: 2024-05-31

Abstract: This thesis explores the documentation practices of digital provenance and paradata within three national libraries: the National Library of Scotland (NLS), the National Library of Spain (BNE), and the National Library of Sweden (KB). By employing a multiple-case study approach, this research examines the types of documentation generated during digitisation projects, the challenges faced by creators, and the implications for the authenticity, reliability, and usability of digital reproductions. The study utilises a qualitative methodology, incorporating data from institutional documents, digital surrogates, and semi-structured interviews with key personnel. The analysis is framed within Bonnie Mak's adaptation of Michel Foucault's archaeological framework, allowing for a critical examination of the socio-technical contexts and decision-making processes that shape digital collections. Findings reveal variations in documentation practices and highlight the need for more standardised frameworks to enhance transparency and trust in digital archives. This research intends to contribute to the broader discourse on digital librarianship and archival science, offering insights that could inform future documentation practices and policy development in digitisation initiatives.

Keywords: Documentation, Digital Provenance, Paradata, Digitisation Process, National Library

"Preserve your memories,
keep them well, what you forget
you can never retell." - **Louisa May Alcott**

"Without libraries, what
have we? We have no past and no
future." - **Ray Bradbury**

Contents

1 Introduction	1
1.1 PROBLEM FORMULATION	2
1.2 AIM AND RESEARCH QUESTIONS.....	2
1.3 CLARIFICATION OF KEY CONCEPTS.....	3
2 Literature Review.....	4
2.1 PROVENANCE AND PARADATA	4
2.2 INTERDISCIPLINARY INSIGHTS AND BROADER APPLICATIONS	6
2.3 USERS AND CREATORS.....	7
2.4 DOCUMENTATION PRACTICES AND CHALLENGES	10
2.5 DOCUMENTATION PRACTICES AND FUTURE DIRECTIONS	14
2.6 SUMMARY	18
3 Theoretical Framework	19
4 Methods.....	22
4.1 METHODOLOGICAL CONSIDERATIONS.....	22
4.2 A MULTIPLE-CASE STUDY	22
4.3 DATA COLLECTION	23
<i>Existing Documents</i>	23
<i>Interviews</i>	24
4.4 THE DATASET	25
<i>The National Library of Scotland (NLS)</i>	25
<i>The National Library of Spain (BNE)</i>	26
<i>The National Library of Sweden (KB)</i>	27
4.5 DATA ANALYSIS	29
<i>Coding Frames</i>	30
4.5 ETHICAL CONSIDERATIONS.....	30
4.6 FINAL REMARKS.....	31
5 Results	33
5.1 DOCUMENTS	33
5.1.1 <i>The National Library of Scotland (NLS)</i>	34
5.1.2 <i>The National Library of Spain (BNE)</i>	36
5.1.3 <i>The National Library of Sweden (KB)</i>	38
5.2 THE INTERVIEWS: INSIDERS' PERSPECTIVES	40
5.2.1 <i>The National Library of Scotland: Digital Gallery</i>	40
5.2.3 <i>The National Library of Spain: Biblioteca Digital Hispánica</i>	44
5.2.3 <i>The National Library of Sweden: Digitala Kollektioner</i>	48
6 Discussion and Conclusions.....	56
6.1 DOCUMENTATION PRACTICES.....	56
6.2 CHALLENGES AND LIMITATIONS	60
6.3 FINAL THOUGHTS AND FUTURE RESEARCH	65
References	68
Appendices	
A: INTERVIEW GUIDE	
B: CODING FRAME FOR THE INTERVIEWS	

1 Introduction

This thesis investigates and describes the documentation practices of three national libraries—the National Library of Scotland (NLS), The National Library of Spain (BNE), and The National Library of Sweden (KB)—with the aim to understand the role that provenance and paradata play in their digitisation processes. It delves into the intricacies of how these libraries document and communicate the digital lineages and creation processes underlying their collections. National libraries were selected for this study due to their critical role within their respective cultural contexts (Lison, 2022). They act as custodians of national heritage, preserving and disseminating the cultural memory of their nations. This includes digitising and providing online access to their collections, a complex challenge as this entails finding a delicate balance between the vast quantity of original materials and the quality of their digital reproductions.

In the digitisation of cultural materials, comprehensive documentation of digital provenance and paradata emerges as a pivotal concern for maintaining the authenticity, integrity and utility of digital archives. This study is particularly concerned with how the histories of original artefacts and their subsequent digital surrogates are interwoven and presented to users, including scholars and the general public. Digital provenance, which relates to the origins and custodial history of artefacts, and paradata, which details the processes involved in creating digital data, are central to determining the authenticity of digital assets. This research will explore documentation practices and how they influence the presentation of digital collections. As libraries and archives transition from physical to digital realms, understanding how digitisation alters or enhances the provenance and paradata of materials becomes essential.

In many ways, documenting digital provenance and paradata is akin to curating a travel journal that maps the journey of a treasured heirloom across generations. Every artefact begins with a story—the moment of its creation, its custodians over time, and how it has evolved. Digital provenance serves as a compass, providing a clear direction to its origins and helping us trace the route taken by each artefact as it changes hands. Meanwhile, paradata is like a set of notes taken during the journey, detailing the landmarks seen along the way—the conversion of a manuscript into a high-quality digital scan, the technological decisions made, and the various hands through which the material passes. Together, these elements create a map that guides researchers, librarians, and users alike through the winding paths leading to the final digital representation.

By examining current practices of documenting digital provenance and paradata, this research also aims to illuminate the often-invisible processes that underlie the creation and curation of digital collections. It seeks to offer insights that could inform future practices in digital librarianship and archival science, thereby contributing to the development of more robust and transparent documentation standards. Through this investigation, we endeavour to understand not only the challenges involved in different documentation

practices but also how the documentation influences the presentation of digital collections as libraries transition from physical to digital realms.

1.1 Problem Formulation

The relationship between a physical object and its digital surrogate is sometimes taken for granted, and the latter seen simply as an unproblematic substitute of *the real thing*. However, the nature of this relationship is much more complex. Moving from the physical to the digital object entails a very nuanced and complicated process that involves several transformations (Björk, 2015). There are issues of trust and authority, authenticity, integrity (Conway, 2010; Dahlström & Hansson, 2019) and usability (Warwick et al., 2009) that arise when creating and interacting with digital surrogates. Quality metadata can provide answers to these issues, strengthening the integrity of a digital object, but some argue that information regarding the object's origin, history and creation should also be documented.

There are several studies that highlight the importance—and argue for—the capturing of digital provenance and paradata when creating digital collections, as we will demonstrate in the literature review, but not much is known about how this works in practice. Furthermore, we have not encountered a large body of literature dealing with the capture and use of digital provenance and paradata (and resulting documents) in library settings. As Dillen (forthcoming) comprehensively outlines, despite the critical role of digital reproductions in the accessibility of cultural heritage, there is a notable gap in the systematic documentation and availability of paradata that explains how these digitisations are carried out. This gap affects the reliability and trust that researchers and the public have in digital archives.

This thesis will attempt to address this gap by exploring three national libraries and their efforts in documenting the digitisation process.

1.2 Aim and Research Questions

This study seeks to explore the purpose and extent of integrating documentation practices for digital provenance and paradata within the digitisation processes of national libraries. Through an examination of three case studies, we aim to address the following research questions:

1. Documentation practices in digitisation projects
 - a. What types of documentation pertaining to digital provenance and paradata are generated in national libraries' digitisation projects?
 - b. Why are these types of documentation generated?
 - c. How do these practices vary across national libraries and projects?
2. Challenges and limitations in documenting provenance and paradata
 - a. What challenges do creators of digital collections encounter when documenting digital provenance and paradata?
 - b. How do these challenges affect the documentation process?

To address the research questions, our thesis employed a multiple-case study approach focusing on the national libraries of Scotland, Spain and Sweden. We collected data through existing institutional documents and semi-structured interviews with key personnel involved in digitisation projects. Qualitative content analysis (QCA) was used to analyse the data, enabling a detailed and nuanced understanding of documentation practices across the three libraries. The results of our analysis will be presented in Chapter 5, while, in Chapter 6, we will discuss the research questions and present our conclusions.

1.3 Clarification of Key Concepts

Before delving into the literature review, it is important for us to clarify some key concepts as we understand them that will be recurrent throughout this thesis. The terms provenance, paradata, metadata, and digital object can have varying interpretations. For the purposes of our research, their meanings are specifically defined as follows:

Provenance: the history of ownership and transmission of an item. In the context of digital collections, it encompasses the documentation of the origins, custody, and changes that a digital object undergoes over time. This concept is vital for establishing the authenticity and reliability of digital reproductions.

Paradata: information about the processes and context surrounding the creation and digitisation of digital objects. This may involve details about the decision-making processes, technical methods, and contextual factors that influenced the digitisation project. Paradata is essential for understanding the conditions under which digital objects were produced and ensuring transparency in digital preservation practices.

Metadata: structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Unlike paradata, which focuses on the context and processes of creation, metadata provides the descriptive, administrative, and technical information about digital objects. Examples include the title, author, creation date, format, and technical specifications of a digital file. Metadata facilitates efficient organisation, discovery, and management of digital resources.

Digital object: also referred to as a digital surrogate, digital reproduction or digital asset is any item that has been digitised from its original physical form or created digitally. This term includes a wide range of materials such as scanned documents, digital images, audio files, and born-digital items. For our research, a digital object is considered in terms of its integrity, usability, and the documentation practices surrounding its creation and maintenance.

Digitisation chain: the series of steps to convert a physical object into a digital format. This should not be confused with image capture, which is just one of the steps. In this thesis, we will sometimes refer to this chain as “the digitisation process” or “creation process”.

2 Literature Review

The exploration is underpinned by a thematic approach to the literature review, which distils significant concepts, practices, and challenges from a wealth of scholarship literature. This methodological choice enables a focused discussion on areas directly relevant to our investigation, providing a foundation for situating our research within the broader academic discourse.

2.1 Provenance and Paradata

In the discourse surrounding the digitisation of cultural heritage and information services, the concepts of provenance and paradata emerge as critical yet complex constructs. Although the significance in the documentation process in digitisation projects has been acknowledged (Ames, 2021; Dahlström & Hansson, 2019; Essen, 2020), a precise delineation between the two remains elusive. This ambiguity stems not from a lack of scholarly attention but from the inherently intertwined nature of these concepts.

Provenance, defined as “the history of ownership of a valued object or work of art or literature”, is traditionally associated with the origin and custody chain of artefacts. The principle of provenance is rooted in the history of archival science, and its formalization is often attributed to the seminal work of Sir Hilary Jenkinson, *A manual of archive administration*, published in the early 20th century. In it, he emphasises the importance of maintaining the original order and context of records through the principles of *le respect pour les fonds* (p. 99) and provenance, advocating for the archivist’s role as a neutral guardian of records. This perspective has shaped modern archival practices and theories.

The principle of provenance has evolved throughout the years. A more recent examination by Tognoli and Guimarães (Merriam-Webster, 2024; 2019), for instance, highlights a contemporary understanding of provenance within archival studies, focusing on the creator, records, and custodial history of archival records. This modern interpretation underscores provenance’s enduring relevance in documenting the history and context of information. At the same time, its scope has also expanded to other fields, as we see in Bearman and Lytle (1985). Building on Jenkinson’s foundation, they explore the potential of this concept beyond traditional archival contexts such as information management practices in organisations. However, despite its central role in archival practices and theory, the concept of provenance has not received as much attention in the fields of libraries, or museums.

The evolution of the principle of provenance, as well as its expansion and current conceptualisation demonstrate its foundational importance and adaptability. This evolution reflects the principle’s capacity to address the ongoing challenges of documenting and preserving information in an increasingly digital world, underscoring its critical role in the knowledge organisation and information management disciplines.

The term paradata was originally used to describe information documenting research processes beyond primary data collection (Cameron et al., 2023). In its traditional interpretation, paradata included both formalized quantitative

information and qualitative observations made during the research process, offering greater transparency and insight into these processes. The concept has evolved over time, gaining a strong foothold in various research fields.

In the library field the term paradata concerns the processual metadata of digitisation. However, there is ambiguity in its definition and purpose of paradata (Sköld et al., 2022). According to Huvila, metadata is data about data (2022, p. 28), whereas paradata encompasses the contextual and processual information about how data has been generated, managed, and manipulated. This concept is important in understanding the lifecycle of data and ensuring its usefulness for future research and applications. Sköld et al. (2022) explore the concept of paradata, emphasising its core essence as information concerning productive processes. This exploration differentiates paradata's focus on detailed process information from the broader spectrum covered by metadata. Moreover, Dahlström and Hansson (2019) refer to paradata as data documenting how digital datasets were collected, curated, and maintained (p. 6), which includes information about decisions made during the digitisation process, the lifecycle and use of the dataset.

When considering the massive amount of data currently being amassed, one possible risk is that it might become ineffectual in the future without adequate knowledge and methods for capturing essential information about the creation, curation, and usage of data. Improving the usefulness of research data starts with better paradata says Huvila (2022).

Ames's article (2021), *Transparency, Provenance and Collections as Data: The National Library of Scotland's Data Foundry*, uses the term provenance in a broad sense that encompasses aspects typically associated with paradata. This could include aspects like documenting OCR, the creation of different file formats, reasoning why and how data is produced, detailing of equipment used, software settings, etc. This reflects an inclusive approach to documenting the lifecycle and transformation processes of digitised collections, which would align with the broader objectives of transparency and trustworthiness in digital archiving practices. This interpretation suggests a nuanced understanding of provenance, acknowledging the importance of detailed process documentation (paradata) in maintaining the integrity and sustainability of digital collections.

The documentation of ancient texts illustrates how digital editions can enhance the transparency and scholarly value of these materials by meticulously recording their provenance and processual metadata (paradata) of these texts to provide comprehensive contextual information (Filosa et al., 2023). This approach not only contextualizes the texts within their historical and editorial landscapes but also sets a precedent for the documentation practices in digital humanities. Researchers can then assess the provenance of a source by examining the documentation provided about its technical and administrative lifecycle. Provenance, in this context, can assess the trustworthiness of digital sources. Since a digital document lacks an original copy in the traditional sense, as it is virtually reconstructed each time it is accessed, establishing its authenticity relies on technical metadata that documents its creation and lifecycle (Essen, 2020). In this view of provenance, we can see an example of

where the functions for provenance and paradata are intertwined. This is of particular importance in traditional print editions which might not have explicitly recorded such information. For digital editions, on the other hand, their technical capabilities, it indicates that it might and arguably “should be standard scholarly practice to include procedural metadata (or “paradata”) in both print and digital editions to contextualise scholarly editions in their historical moment” (Filosa et al., 2023, p. 60). Such an inclusion would not only contextualize scholarly editions in their historical moment but also acknowledges the change technological and scholarly landscape in which they operate.

As we have seen, the perspectives on paradata and provenance vary across different articles and disciplines, reflecting the nuanced understanding of these concepts in the context of digital documentation and archival practices. While some sources may treat paradata as distinct from provenance, emphasising its role in documenting the process and methodology of data creation, others may see paradata as encompassing provenance, or even as a specific form of provenance, highlighting the intertwined nature of data’s origin, context, and the methodological details of its generation. This diversity in interpretation underscores the complexity of ensuring data integrity, authenticity, and reusability in the digital era. Understanding how paradata is conceptualized, created, and used can inform us on documenting and preserving digital provenance and paradata in digitising institutions. The use of paradata for enhancing transparency, research robustness, data evaluation, and cross-disciplinary communication can be particularly valuable for understanding the dynamics of digital archiving and library science.

2.2 Interdisciplinary insights and broader applications

The digitisation of cultural heritage and scholarly content marks a significant evolution in the preservation and dissemination practices within libraries. This literature review, while primarily centred on libraries, extends its inquiry to encompass insights from the wider domain of cultural heritage, and other specialised fields like archaeology (Börjesson et al., 2020, 2022), archives (Cameron et al., 2023) and survey research (Couper, 2017; Kreuter, 2018), incorporating a diverse array of scholarly contributions, which reveal the evolving practices and challenges within these various fields. Doing this will allow us to critically evaluate the practices surrounding the documentation of digital provenance and paradata during the digitisation process.

Lemieux and the imProvenance Group (2016) explore the concept of provenance across various disciplines, highlighting its evolving understanding and the challenges of capturing, representing, and using provenance information in increasingly distributed and heterogeneous information ecosystems. This synthesis draws from archival science, law, computer science, library and information science, and visual analytics to offer a multidimensional view of provenance, emphasising its role in establishing trust, authenticity, and the reliability of data and documents, as well as in supporting decision making processes. Provenance, traditionally a core principle in archival discourse, as we have mentioned earlier, has gained

attention in other fields due to the growing use of information and communication technologies.

Significant emphasis is placed on the challenges of documenting provenance in digital environments, where the integrity of digital objects must be preserved across technological changes and potential obsolescence. Lemieux and the imProvenance Group advocate (2016) for the application of sustainable development principles to the management of digital cultural heritage, which involves strategies that ensure the long-term preservation, accessibility, and usability of digital resources. Such strategies are designed to meet current needs while safeguarding future generations' ability to do the same. They particularly emphasise the importance of thoroughly documenting contextual information (paradata) to mitigate epistemic failures caused by rapid technological advancements and the subsequent obsolescence risks. The concept of "epistemic failure", as also discussed by McCarthy (2007), presents significant challenges; this will be further explored in the 'Documentation Practices and Challenges in Digitisation Projects' section further down on our literature review.

The discussion extends to methods for capturing and representing provenance, ranging from archival arrangement and description practices to advanced computational approaches in e-science and visual analytics. These methods illustrate a shift towards leveraging technology for provenance documentation, though challenges remain in standardising and sharing provenance information across disciplines and systems. Kreuter's work (2018) in the context of survey research offers valuable insights that can be applied to the field of digital libraries, particularly concerning the documentation and communication of digital provenance and paradata. She discusses the use of paradata to enhance efficiency in survey data collection, identify errors, and adapt collection strategies in real-time.

Drawing on Kreuter's detailed exploration of paradata within survey research, several key aspects emerge that are pertinent to understanding and enhancing documentation practices in digital libraries. These aspects include the strategic utilization of paradata for operational efficiencies, the dynamic adaptation of data collection methods in real-time, challenges presented by the transfer of data across different platforms and institutions, and the delicate balance between customisation and standardisation of data practices. Each of these points provides a valuable framework for examining the similar challenges and opportunities present in the documentation of digital provenance and paradata within digitisation projects.

By examining provenance and paradata from an interdisciplinary perspective, this section underscores its complexity as a dynamic construct and the importance of cross-disciplinary collaboration to enrich and expand traditional definitions and applications.

2.3 Users and Creators

Much of the literature highlights the importance of documenting the digitisation process. However, it is worthwhile to look at who are the creators

of this documentation and the users of the digital reproduction. Paradata creators are usually researchers or professionals involved in the creation of the digital product, while users include a broad range of stakeholders: experts, non-experts, professionals, and the public. The potential application of paradata extends beyond academic research, suggesting its usefulness in other domains like cultural heritage and public engagement. Sköld et al. (2022) identify three main modes of paradata use:

- **Research Robustness:** It enhances the methodological rigour and reliability of studies, communicating key components of scholarly production and underpinning hypotheses and interpretations.
- **Data Evaluation and Assessment:** Paradata facilitates the evaluation and assessment of research data, improving its reusability and sustainability.
- **Cross-Boundary Communication:** Paradata can aid in overcoming communication challenges between diverse research and user groups, making scholarly processes more visible and understandable to those from other domains.

Documenting information processes and practices is a complex and multifaceted task, influenced by the context in which it occurs (Huvila et al., 2021). Paradata must be gathered from multiple sources. To produce useful paradata, it is necessary to recognize its value and incorporate its capturing and preservation into daily practices. Combining different sources of paradata is essential, especially for digital data. The complexity of defining what constitutes paradata and determining how it can be captured and used effectively is a significant challenge (Sköld et al., 2022). A critical aspect is balancing the amount of paradata being captured. Huvila et al. (2021) highlight that this balance is critical for making documentation comprehensive yet manageable, ensuring it is useful without being overwhelming. The aim is to capture enough detail to provide context and support the integrity and reuse of digital collections, while considering practical constraints like resources and the documentation's relevance to users. Essentially, this means documenting processes in a way that is detailed enough to be informative but streamlined enough to be practical and accessible.

Another aspect of documentation highlighted in the literature is recording when an object changes location. With regards to provenance in digitisation, recording the detailed location-based information of objects should be an integral part of their digital provenance (Padfield et al., 2019). This would not only include their physical history but also the digital journey that these objects undertake as they are digitised and made available online. Such an approach would not only be relevant to museum objects, but also to digital collections in libraries and archives. Understanding where an object has been and how it has been digitised and presented online is important for authenticity and scholarly research. Once more, documenting provenance and paradata in the context of digitising and managing cultural heritage collections, reflects the need for detailed record-keeping, standardisation, and the use of advanced technologies.

The papers by Hirtle (2002) and Conway (2010) seemingly present contrasting perspectives on the effect of digitisation on the use of original material in special collections. The discrepancies underscore the complexity of the

digitisation's impact and suggest the consequences can vary depending on a range of factors, including the nature of the collections, the goals of digitisation, and behaviours of users. Hirtle suggests that users may prefer accessing material digitally rather than physically, "if you make special collections material available via the Web with appropriate metadata [...] they will be used" (p. 43), where Conway sees a potential increase interest in physical collections due to the transformative effects of digitisation, particularly for purposes such as research. The impact of digitisation on the use of original materials in special collections is nuanced. It can lead to both a potential decrease in physical handling for routine access needs and an increase in interest and engagement with physical collections for specialised research or when the digitisation effort highlights unique aspects of the collections. This dual effect illustrates the ongoing need for balanced strategies in managing and promoting both digital and physical collections.

On a not dissimilar line, Björk (2015) discusses this inherent paradox in preserving text-based documents: the need to minimize their usage to extend their lifespan, versus the necessity of their transmission and movement to ensure survival and accessibility. He argues that the status of digital documents varies depending on institutional contexts and knowledge organization frameworks, influencing how these documents are treated and perceived, therefore the digitisation processes and the digital transformation of text-based documents can impact their provenance and the metadata (or paradata) associated with these processes.

Now, focusing on the users, and how they perceive digital resources when they access them, the work by Warwick et al. (2009) centres on the importance of documentation for digital humanities resources. This includes both technical documentation of textual markup or database construction and procedural documentation about resource construction, which reflects the essence of paradata. They conclude that good documentation is essential for the successful use and reuse of digital resources in the humanities. They also advocate for standardisation of how those who create projects document their work. The availability and quality of documentation impact user perceptions of resource quality and usability further underline the significance of paradata. Warwick et al. indicates that users seek information that assures them about the resource's construction, quality, and suitability for their research needs, which is precisely what well-structured metadata aims to provide.

Huwe (2023) stresses the importance of metadata in driving long-term strategies for digital library development, including user education and engagement. The article underscores that the real work in building digital collections lies in developing and managing metadata to facilitate better access and discovery for users. Dillen (forthcoming), in the context of scholarly editing, also discusses the importance of providing comprehensive paradata to help users from diverse backgrounds assess the quality and provenance of digitised materials. The availability of detailed paradata is crucial for users to understand the digitisation process and trust the digital reproductions. Proper documentation of digitisation processes, including paradata, helps ensure the reliability and usability of digital reproductions, however there are some

challenges, while detailed paradata is beneficial, it is often difficult to obtain and document comprehensively. The document discusses the need for a balanced approach that provides enough paradata to be useful without overwhelming users with excessive detail. This sentiment echoes Huvila et al. (2021) as mentioned above.

2.4 Documentation Practices and Challenges

There are challenges in documenting intellectual processes in research (Huvila, 2022) as well as documenting and describing processes and practices in the field of information science and technology (Huvila et al., 2021). Past efforts have mainly focused on conceptual and the technical aspects of paradata, without in-depth exploration of what should be documented and how.

The application of paradata in the digitisation process can document the context of a dataset's creation, development, and maintenance (Dahlström & Hansson, 2019, p. 6). This would include decisions made during digitisation, the history of the digital provenance, and the maintenance life cycle. Also, paradata can enhance transparency and authenticity, as it helps ascertain the authenticity and usability of digital reproductions by providing information about the production and conversion processes. However, Dahlström & Hansson indicate that the implementation of paradata can present some challenges, paradata is not always present or is limited, plus outsourcing steps in the digitisation process can make it harder to track the history of how the digital versions were made and impact our ability to do so. As a solution the paper suggests that paradata could be crucial in addressing issues related to version history, authenticity, and usability of digital collections. It can assist researchers in understanding the nature of digital reproductions and in making informed decisions about their use. However, at this point, there is no established best practice with regards to documenting paradata (for digitisation processes) yet, especially for image materials. Some projects document paradata in project reports, about-pages, or internal wikis, but there is no general consensus about what could or should be included. Dillen (forthcoming) provides a detailed account of the negotiation and variability in quality standards during the digitisation of manuscripts, illustrating the complex challenges institutions face in maintaining consistent documentation of digital provenance and paradata.

Some empirical studies provide insight into the practical challenges and innovations in documenting digital provenance and paradata. In the paper "Archaeology of a Digitization", Mak (2014) engages with the concept of documentation, albeit through a more theoretical lens, focusing on the processes and implications of digitisation rather than explicit procedural aspects like documentation practices. The author argues that the practices of digitisation, from transcription to selection criteria, not only contribute to the digital representation of historical texts but also influence the construction of knowledge and the authoritative status of these digital artefacts. Mak's study critically examines the Early English Books Online (EEBO) database, highlighting the ontological differences between digital reproductions and their original artefacts. She argues that these differences have not been fully acknowledged or understood within the digital humanities and library sciences.

By analysing EEBO, Mak reveals how digital surrogates are not merely technical reproductions but are imbued with current perceptions of the past, thus influencing scholarly research and the production of cultural heritage. The document sheds light on how digital projects can shape the understanding and accessibility of historical literature, raising important questions about authority, authenticity, and the future of scholarly research in the digital age. She suggests that digitisations should be analysed as “material, bibliographical object” (2014, p. 1515) that carry clues about their creation and dissemination processes.

The study calls for more critical and systematic approaches to documenting the creation and circulation of digitised materials, emphasising the need for transparency and accountability in digital library practices. We will revisit this study in Chapter 3, where we will discuss the theoretical framework that will guide our thesis.

Another relevant empirical study by Jarlbrik and Snickars (2017) delves into the multifaceted issues surrounding the digitisation of historical newspapers in Sweden, emphasising the interpretative challenges posed by optical character recognition (OCR)¹ technology. This scholarly work navigates the technical, historical, and cultural dimensions of digitising archival materials, specifically focusing on nineteenth-century newspapers and how digitisation processes translate these physical documents into digital formats. The authors begin by contextualising their study within the broader discourse of digital humanities and media studies, highlighting the transformative impact of digitisation on archival research and the accessibility of historical documents.

They draw upon concepts from communication and noise to frame their analysis of digitised newspapers as instances of cultural heritage transformed into “digital noise.” Their paper presents a case study to illustrate the practical challenges and considerations involved in digitising historical newspapers. These examples highlight the technical complexities of OCR, the importance of metadata, and the interpretive nuances required to navigate digitised archives effectively. Their findings reveal significant amounts of noise, including millions of misinterpreted words and texts incorrectly edited or generated by the digitisation tools. This noise, they argue, arises from the outsourcing of digitisation processes and the use of commercial software packages, which are often black boxes to library staff, leading to a loss of control over the quality and provenance of digital collections.

The paper highlights several key points that seem relevant:

- Digitisation transforms newspapers into new objects, raising questions about the authenticity and utility of digital reproductions.
- Libraries risk losing control over their collections when digitisation processes are outsourced or rely heavily on commercial software.

¹ OCR: Optical Character Recognition – technology that converts different types of text, such as scanned paper documents or images of text, into machine-readable and searchable digital formats.

- The software used in digitisation (like OCR and auto-segmentation tools²) significantly impacts the quality of digital collections, often introducing errors and "noise" that can distort the historical record.
- The authors conducted a limited ethnographic study of the digitisation process, revealing a lack of knowledge among staff about how digitisation tools work and their implications for digital collections.
- There's a need for better understanding and oversight of digitisation processes to ensure the preservation of cultural heritage in the digital realm.

Jarlbrik and Snickars (2017) conclude that while digitisation is crucial for preserving newspapers and making them accessible, the process is not neutral and significantly alters the documents. This transformation challenges the notion of provenance and raises important questions about the nature of digital archives and the reliability of digitised historical documents. They argue that digitisation is not merely a technical process but also a cultural and interpretative act that shapes how historical materials are accessed, understood, and utilized in contemporary research.

Gravan McCarthy (2007) discusses the importance of managing comprehensive information to meet the varied needs within the cultural heritage community. He acknowledges the challenges in ensuring the long-term viability of digital cultural heritage resources, particularly amid rapid technological changes and potential obsolescence. The author identifies two major challenges: media redundancy due to technological evolution and epistemic failure. The latter, a concept previously introduced, refers to the loss, misinterpretation, or inaccessibility of information due to the absence or inadequacy of contextual knowledge surrounding it. This includes preserving information about the provenance and significance of digital objects in a manner comprehensible to those outside the immediate heritage community, as well as to future generations.

McCarthy (2007) highlights the limitations of traditional catalogue systems and underscores the need for more comprehensive frameworks that include broader contextual information. McCarthy argues for the systematic preservation of contextual information as a means to mitigate epistemic failure. He highlights the inadequacy of existing systems and practices for long-term information preservation, emphasising the need for a more effective transfer of digital objects and information to future generations. The chapter suggests that a network approach, which includes detailed documentation of contextual information such as the creators, the intra- and interorganisational relationships, and the standards framework, could be more effective than traditional cataloguing methods (keeping in mind that this was written in 2007).

² Auto/segmentation tools are automated technologies used in digitisation to divide and annotate text, images, audio, and video, enhancing efficiency and accuracy. They employ techniques like OCR for text and speech recognition for audio, though their effectiveness varies with content quality and complexity.

The inevitable loss of contextual information over time, particularly in inadequately managed systems, is identified as a threat to effective knowledge transfer. The document calls for recognition of specialist knowledge and strategies to prevent its loss, highlighting the critical role of comprehensive documentation practices in the sustainability of digital cultural heritage. The challenge is working with incomplete or varied metadata formats in digitised collections, therefore effective organisation, interpretation, and accessibility of digitised collections largely depends on well-structured metadata (paradata) (Huwe, 2023).

The movement of skilled and knowledgeable personnel, leading to the loss of implicit or contextual information, exacerbates this issue. McCarthy (2007) proposes that a contextual information framework, which maps the relationships between different entities involved in the creation and management of digital cultural heritage, could provide a robust structure for preserving knowledge over time. In 2007, McCarthy envisioned frameworks as electronic networks that could link together sources of knowledge, potentially via the World Wide Web—a vision that aligns closely with contemporary practices in information sharing. His foresight anticipated the digital infrastructure that now underpins the preservation and accessibility of vast amounts of information and knowledge involved in the creation and management of digital cultural heritage resources. This approach was designed to ensure that future generations could comprehend and confidently use this information. Reflecting on McCarthy’s vision from today’s perspective, it is evident that many aspects of his early predictions have been realised, highlighting the enduring relevance of his contributions to the field.

As our thesis delves into the documentation practices of digital provenance and paradata in libraries, we now focus on the empirical insights from Symonds and May’s (2009) study on the “Dear Comrade Project”. This research provides a critical empirical exploration of standard digitisation processes within library settings. The authors detail the procedural framework they developed for the digitisation of the Eugene V. Debs correspondence collection, focusing on structured project management and controlled vocabulary. Key practices such as meticulous planning, systematic organization and naming of files, detailed metadata creation, and rigorous quality control are emphasized. These measures not only enhance the usability and reliability of digital reproductions but also ensure the authenticity of digitised collections. By presenting practical experiences, observations, and outcomes from implementing these processes, the paper contributes significantly to our understanding of the challenges and strategies in managing digital assets. This empirical study offers valuable lessons and replicable models for similar digitisation efforts, which are crucial for understanding the intricacies of digital provenance and the impact of paradata in library digitisation projects.

Lucie C. Burgess (2016) delves into the multifaceted notion of provenance within digital libraries, as observed in the Bodleian Digital Library at the University of Oxford. Provenance, in this context, extends beyond traditional notions of authorship or origin to include aspects like information integrity,

data rights, and digital content exploitation. The document highlights the evolutionary nature of digital libraries, emphasising how they adapt and transform to accommodate new research outputs and user needs. It examines the application of provenance through various library projects, underlining its role in enhancing information quality, enabling effective discovery, and fostering trust in digital resources. Challenges in capturing structured, machine-readable provenance metadata from analogue content are also addressed, pointing out the need for extensive research, digitisation efforts, and the development of innovative methodologies to manage and interpret provenance information effectively.

2.5 Documentation practices and Future Directions

Dahlström and Hansson (2019) have laid the groundwork of the complexities introduced by digitisation, pointing out that digital reproductions are not mere facsimiles of original documents but are products of intricate processes that merit thorough documentation. They discuss the relationship between digital reproductions and their physical source documents, emphasising the complexities introduced by digitisation processes. They argue for a critical perspective on digitisation, noting that digital reproductions are not just straightforward copies of original documents. The digitisation process, including pre-processing, processing (like image and text capture), post-processing (metadata editing, versioning), and long-term maintenance, involves various professional decisions that impact the final digital product. The paper highlights several concerns related to digital reproductions. Firstly, it notes a tendency to accept digital reproductions without sufficient scrutiny. Secondly, it discusses how outsourcing digital reproduction processes can complicate the tracing of their provenance, and lastly, it addresses the risk of altering digital images in ways that could undermine their authenticity and reliability. These issues collectively pose significant challenges in maintaining the integrity of digital reproductions.

Whearty's (2023) *Digital codicology: Medieval books and modern labour* delves into the early stages of manuscript digitisation, comparing them to the incunabula period of printing to highlight the experimental and formative nature of these endeavours. Her book underscores the importance of metadata in making digital manuscripts discoverable and usable, highlighting how the history of digitisation has been marked by projects that, while successful in isolation, have struggled to integrate with broader digital ecosystems due to "siloes" data (2023, p. 168).

She narrates the initial challenges faced by digitisation projects, such as technological limitations, lack of standards, and questions about best practices for digital reproduction. Whearty (2023) argues that similar to how incunabula printers experimented with typography, layout, and distribution, early digitisation projects navigated uncharted territories to establish methodologies and conventions for digital manuscript reproduction. She challenges the notion that digitisation is a neutral process, highlighting how it inherently involves choices that shape the visibility and accessibility of cultural heritage. Whearty advocates for the documentation of the digitisation process, including decisions made and challenges encountered. This documentation can serve as a resource

for understanding and contextualising glitches (material traces of digital processes and human labour), as well as for guiding future digitisation efforts to avoid similar issues.

Hardy (2018) also addresses the acknowledgment of human labour in the digitisation process. Hardy's call for a re-evaluation of the recognition given to human labour emphasizes that the individuals behind the digitisation process are invaluable assets to the archival field. This labour, often overshadowed by the digital outcomes it produces, requires greater visibility and appreciation within the academic narrative. By integrating these insights, Hardy proposes that our approach to digital archives and scholarship should reflect the integral role that human intervention plays. The juxtaposition of Whearty's and Hardy's viewpoints underscores a common theme in the literature: the intricacies of the digitisation process are profoundly shaped by the interplay between human labour and technology.

Essen (2020) illuminates the role of documentation, particularly provenance, as historical research transitions into the digital era. This concept underscores the important role that the origins and lineage of digital artefacts play in scholarly endeavours. By tracing the journey of digital objects from their physical counterparts through to their current digital existence, Essen argues for a robust and transparent provenance practice. Such meticulous record-keeping not only reinforces the authenticity and integrity of historical digital collections but also ensures their meaningful use in research and analysis. Essen's discussion on the indispensability of technical metadata in establishing the authenticity of digital documents, aligned with the Open Archival Information System (OAIS)³ model, reinforces the imperative for comprehensive documentation standards. The adoption of this model offers a structured approach to digital preservation, emphasising the importance of metadata in documenting the provenance, context, and access rights of digital objects. This framework is pivotal in ensuring the long-term accessibility and usability of digital information.

The OAIS model reminds us of the FAIR principles (Wilkinson et al., 2016), which seek to make data: findable, accessible, interoperable and reusable. These guidelines seek to enhance the management and sharing of data in the digital age across various disciplines, fostering collaboration and efficiency in research. However, not everyone views the FAIR principles as a working framework. Dunning et al. (2017), for instance, delve into the effectiveness and applicability of the FAIR data principles by evaluating data archives' adherence to these principles. Their analysis indicates significant challenges in the practical implementation of the FAIR principles, particularly with the aspects of interoperability and reusability. The findings suggest that less than half of the repositories analysed were fully compliant with the FAIR guidelines, highlighting a gap between the principles' theoretical ideals and their practical application. The study concludes that while the FAIR principles

³ OAIS: Open Archival Information System – a conceptual framework that provides guidelines and standards for the long-term preservation and management of digital information, ensuring the integrity and accessibility of archived data over time.

provide a valuable framework for improving data sharing and reuse, their application is not straightforward. Compliance is not merely a matter of checking off requirements but involves interpreting and adapting the principles to fit the diverse contexts of data repositories.

The FAIR principles are designed to enhance research data management and stewardship. It is important to note that these principles are primarily focused on technical standards that facilitate computer-to-computer interactions, such as through APIs, rather than human-computer interactions via graphical user interfaces. For example, the ‘Accessibility’ principle in FAIR emphasizes standardising technical methods for accessing digital resources, such as data retrieval protocols accessible through computational means, rather than focusing on user interface design or web accessibility for individuals, for instance, with disabilities. This distinction is essential as adhering to FAIR principles establishes a technical groundwork for data management that can later support the creation of applications aimed at improving user interaction and inclusivity. Consequently, while FAIR principles play a significant role in promoting the technical interoperability and reusability of digital research data, they do not cover all aspects of accessibility, highlighting the need for additional strategies to address the full range of data accessibility issues.

The introduction of innovative methodologies such as provenance visualisation (Hart & de Vries, 2017), collections as data (Ames, 2021; Padilla, 2017; Padilla & Higgins, 2014) or data collections upcycling (Scheltjens, 2023) augment the transparency and interpretability of digitised collections. Such transparency and explicitness in documenting the digital editing process, including decisions made, sources used, and the methodologies applied, cater to a more inclusive, sustainable, and scholarly practice. It facilitates a deeper understanding of the material and immaterial aspects of ancient texts and objects, ensuring that digital editions and collections can be a reliable and rich resource for both current and future scholarship. These methodologies underscore the potential of data visualisation techniques in making the provenance of records interactive and accessible, thereby fostering a deeper understanding of the records’ history and the labour involved in their transformation. However, Hart and de Vries (2017) also advocate for human intervention without negating the value of automation in metadata creation. Some aspects such as descriptive and administrative details require human judgement and intervention to accurately capture the context, history, and provenance of digital objects. Hart and de Vries’ approach aims to enhance the transparency and interpretability of historical records by documenting and visualising the changes and decisions made throughout their digitisation and curation processes. It leverages data visualisation techniques to make the provenance of records interactive and accessible, thereby promoting a deeper understanding of the records’ history and the labour involved in their transformation. This method has the potential to enrich research by allowing scholars to explore and interrogate historical datasets in novel ways, encouraging ethical considerations and more rigorous historical document analysis.

The concept of “collections as data”, advocated by Padilla (2017), further broadens the discourse, urging libraries to reconceptualize digital collections as computationally amenable data. This paradigm shift emphasizes the need for structured data practices that enhance the legibility and generativity of digital collections, thereby promoting their use in computational research and fostering a culture of open, transparent data practices. In Padilla’s rationale, the concept of legibility emphasizes the need for digital collections to be understandable not just by humans but also by machines, making data accessible and interpretable through proper documentation and metadata. This enhances the ability of both researchers and algorithms to engage with digital archives meaningfully, promoting the use of collections in innovative computational research. Furthermore, the concept of legibility translates also to better data, in our case: metadata, provenance, paradata, etc. As such, Padilla’s concept refers to the idea of enhancing the capacity for meaning-making by making collections more amenable to a broad set of uses, particularly through computational methods. This would involve rethinking the form, description, discovery, and access of collections to expand their usefulness beyond traditional research methods. Generativity on the other hand, involves fostering an environment where libraries and their collections can support a wide array of inquiries and explorations, encouraging innovative and diverse approaches to understanding and interpreting data. This concept underlines the importance of not just preserving and providing access to collections but actively enhancing their potential to generate new knowledge and insights.

The importance of provenance in making collections data usable and trustworthy is explored by Ames (2021). Her study discusses challenges and strategies for documenting the transformation of collections into data, ensuring open access, and maintaining data integrity. This work aligns with broader trends in digital scholarship and the GLAM⁴ sector toward open, transparent data practices. Following this line of thought, a novel approach is introduced to enhance the scientific value of historical data collections (HDC) through upcycling (Scheltjens, 2023). Here, upcycling, is conceptualised as a research practice aimed at increasing the usability and interoperability of HDCs, focusing on the reuse and integration of data, and examining historical information creation processes. Upcycling, according to Scheltjens (2023), emphasises the importance of documenting and analysing paradata—the data about the data processing activities—to understand the cognitive processes involved in historical research. By applying the above mentioned FAIR principles to HDCs, Scheltjens advocates for a paradigm shift in digital history that appreciates the value of past research efforts and leverages them for future inquiries.

In the ongoing project “*Documentation of data making, processing, and use facilitates future reuse of research data: the CAPTURE project*,” Huvila and Ekman (2024) focus on the importance of paradata in research data management. The CAPTURE project, funded by the European Research Council, investigates what information about the creation, management, and

⁴ Stands for Galleries, Libraries, Archives and Museums, and refers to the collective institutions that are dedicated to preserving and conserving cultural heritage.

use of research data is necessary for its future reusability. It also explores efficient and comprehensive methods to capture this information. The project emphasizes the balance between documenting enough information without being overly labour-intensive and the challenges of standardisation in data management. The empirical focus of CAPTURE is on archaeology, a field that deals with a wide range of data types, making it an ideal context to study data documentation complexities. The project aims to contribute to standards and tools for paradata, supporting national and global policies for data management and open data. The project focus on standardisation can be a challenge “considering the significant extent to which paradata is coincidental and exists because of the lack of data cleaning and management, a major challenge is also how to strike a balance between too much and too little standardisation” (Huvila & Ekman, 2024, p. 28).

2.6 Summary

The literature review has explored the multifaceted domain of documentation practices in library digitisation projects, with a specific focus on the generation and management of digital provenance and paradata. It has revealed a consensus on the importance of these practices for the authenticity, reliability, and usability of digital reproductions. Despite this agreement, the review also uncovered a diversity in practice and a series of challenges, including varying institutional standards and the evolving nature of digital technologies. Key scholarly contributions have provided foundational insights into the operationalisation of paradata and the nuanced understanding of digital provenance. This body of work has underscored the necessity for a concerted effort towards the development of more standardised and comprehensive documentation frameworks. As our research proceeds, it will draw on these insights to address identified gaps and contribute to the enhancement of documentation practices within the realm of digital libraries.

3 Theoretical Framework

In the evolving landscape of digital librarianship, the documentation of digital provenance and paradata emerges as a critical area of inquiry. This chapter delves into the theoretical framework that underpins our exploration of documentation practices within three national libraries. Bonnie Mak's (2014) innovative adaptation of Michel Foucault's (1972) archaeological framework provides a compelling lens through which to scrutinize the documentation practices of digital libraries. Mak's scholarly inquiry into digitisation processes not only challenges conventional perceptions of digital objects but also foregrounds the intricate layers of decision-making and cultural practices that shape them.

Michel Foucault's *The archaeology of knowledge* presents a radical shift from traditional historical and analytical methods. Foucault proposes archaeology as a metaphor for uncovering the rules, systems, and discursive formations that underlie the production of knowledge within specific historical contexts. Central to Foucault's archaeology is the concept of the discursive formation – the body of rules that dictate what can be said, by whom, and under what conditions. This framework challenges the continuity and progress narrative, focusing instead on breaks, discontinuities, and the complex interplay of power and knowledge. Foucault's emphasis on the statement as a function of existence rather than as a carrier of meaning introduces a new way of understanding the materiality of discourse and its role in shaping knowledge.

Foucault's archaeological method provides a robust lens through which to examine the practices surrounding digital reproductions in libraries. By adopting this framework, we can explore how digitisation projects are not merely technical endeavours but are deeply embedded in the institutional, social, and cultural contexts that govern their execution. This approach allows us to question the assumptions and values that underpin decisions in the digitisation process, from selection criteria to metadata standards, and to consider how these choices shape the digital objects that emerge from these processes.

Bonnie Mak's examination of digitisation through Foucault's archaeological method illuminates the nuanced processes underlying the creation and curation of digital collections. This approach draws attention to the materiality of digital objects, emphasising that digitised artefacts are not mere replicas of their physical counterparts but are instead reconstituted through a complex interplay of technological, institutional, and social factors.

- Contextualising digital objects: Mak's methodology urges us to consider digital objects within their broader socio-technical systems. By treating digitised materials as "material bibliographical objects," she invites a consideration of how digitisation practices reflect and reinforce specific values and priorities within the library and archival communities. This perspective allows us to question which aspects of the original materials are emphasized or obscured in the digital realm and how this shapes users' engagement with digitised collections.

- Documenting the invisible: A key insight from Mak's approach is the critical role of documentation in making visible the often-invisible processes of digitisation. This encompasses not only the technical aspects of scanning and metadata creation but also the selection, interpretation, and representation of choices that inform the digitisation process. By applying an archaeological lens, we can unearth the layers of decision-making that contribute to the construction of digital objects, thereby enhancing our understanding of digital provenance and paradata.
- Reconsidering authority and authenticity: Mak's framework challenges us to rethink notions of authority and authenticity in the digital domain. The processes by which digital objects are created, annotated, and disseminated are laden with choices that carry implications for the perceived authenticity and scholarly value of digitised materials. Understanding these processes through an archaeological approach underscores the importance of transparent documentation practices that articulate the provenance and paradata of digital objects.

This approach allows us to question the assumptions and values that underpin decisions in the digitisation process, from selection criteria to metadata standards, and to consider how these choices shape the digital objects that emerge from these processes.

Adopting Mak's archaeological approach enables us to critically examine the documentation practices of digital provenance and paradata within libraries. This perspective encourages a reflection on the practices that govern the creation of digital artefacts, urging a consideration of the ways in which libraries can document the complex processes of digitisation to ensure the authenticity, reliability, and usability of digital reproductions. It prompts us to think about documentation not just as a technical requirement but as a critical component of digital stewardship that contributes to the scholarly and cultural value of digital collections. Such documentation should not only record technical parameters and processes but also the rationale behind selection decisions, the criteria for metadata standards, and the intended audience and use cases for digital collections.

Drawing from Mak's insights, we advocate for a more holistic approach to documentation that encompasses both provenance and paradata. This entails developing documentation practices that are capable of capturing the multi-layered nature of digitisation projects, from the technical to the interpretative. By doing so, libraries can foster greater transparency, enabling users to critically engage with digital collections and researchers to better understand the context of their creation.

The archaeological framework, as outlined by Foucault and applied by Mak to digitisation, offers a comprehensive theoretical basis for examining documentation practices of digital provenance and paradata in libraries. By foregrounding the materiality and constructedness of digital objects, this approach calls for a re-evaluation of what is documented and how, ultimately aiming to cultivate a richer, more nuanced understanding of digital collections.

This perspective facilitates an understanding of the discursive and material conditions shaping digital collections and underscores the importance of transparent, reflective, and informed documentation practices.

In this thesis, we will apply this theoretical framework, specifically Foucault's Archaeology and Bonnie Mak's archaeological approach to digitisation, to our case studies of the National Libraries of Scotland, Spain, and Sweden (Chapter 6, Discussion & conclusions). By examining the documentation practices of digital provenance and paradata through this lens, we aim to uncover the discursive formations, material conditions, and decision-making processes that shape digital collections. This approach will guide our analysis of the collected data, allowing us to critically evaluate how each library constructs and communicates the digital lineage of its collections. This critical evaluation will enhance our understanding of the authenticity, reliability, and usability of digital reproductions. Part of this "archaeological work" involves unearthing the motivations and challenges underlying the documentation practices. While documents provide one part of the story, the voices of those involved complete the picture and offer essential explanations.

4 Methods

4.1 Methodological Considerations

This thesis presents a descriptive and exploratory study, as it is appropriate if the purpose is to gain a deeper understanding of a phenomenon or setting that is “too complex to take in in just a superficial observation”, unknown or not fully defined (Wildemuth, 2017a, p. 28). In our case, we seek to explore how the provenance of digital objects and collections, as well as the paradata associated with their creation, are documented in a specific type of setting: national libraries.

Descriptive and exploratory studies can take either a quantitative or qualitative approach. The nature of our research problem and questions has led us to adopt a qualitative stance, where we have worked inductively, using the data to draw insights and conclusions. After all, the aim is not quantifying a phenomenon; rather, we are interested in producing a detailed description of what is currently taking place in the three selected national libraries.

4.2 A Multiple-Case Study

Case studies are a “detailed and intensive analysis” of one or more cases (Bryman, 2016). Adopting this approach offers several advantages: first, it is ideal for descriptive and exploratory analysis (Choemprayong & Wildemuth, 2017), as zeroing in on just a few cases allows researchers to gather a large amount of data that can later be used to paint a vivid picture of a phenomenon or setting. The second advantage lies in its flexibility, as different methods of data collection and analysis can be used and combined (Choemprayong & Wildemuth, 2017). Thus, a research problem can be studied from multiple angles and perspectives.

Our research takes the form of a multiple—or comparative—case study and focuses, as mentioned above, on three European libraries: the National Library of Scotland (NLS), the National Library of Spain (BNE), and the National Library of Sweden (KB). Choemprayong and Wildemuth (2017, p.54) explain that comparative case studies contribute “to cross-case analysis and the extension of theory to additional individual or settings” and one way of selecting the cases is following a logic of literal replication. We followed this logic to some extent, in the sense that the three selected cases have aligned missions—including fostering research⁵—and objectives regarding digitisation. At the same time, there are still differences in how they implement their digital strategies, which yields interesting comparisons.

⁵ In “Policy för digitisering av Kungliga bibliotekets samlingar” (2022; see KB2), the KB states: Part of the mission of the library is to provide source material to foster research and democratic social development” (p. 3; our translation); In “The Library’s approach to selection for digitisation”, the NLS states: “The resulting digital resources will promote opportunities to advance learning and, foster and develop new research” (see NLS3); and, finally, in relation to Biblioteca Digital Hispánica, BNE states that BDH aims at becoming a key resource that fosters research on Spanish culture (Biblioteca Digital Hispánica; our translation; see BNE4).

Regarding our case selection, from the very beginning we wanted to work with representative or typical cases. Bryman (2016, p. 62) calls these “exemplifying cases” and states that they “are often chosen not because they are extreme or unusual (...) but because either they epitomize a broader category of cases or they will provide a suitable context for certain research questions to be answered”. This is well-suited to this research as our impression is that our topic has not been fully explored, so before trying to describe the exception to the rule, we wanted to understand the standard.

Nevertheless, it is important to note that ours is a convenience sample as both our current geographical location, as well as our personal and linguistic backgrounds, allowed us access to these European libraries. The logic of selecting the KB was that we both live and study in Sweden and are engaged in the study of library and information science within this context. This familiarity and interest in the library’s digital initiatives, combined with guidance and assistance from our supervisor in connecting with study participants, made the KB an appropriate choice.

The inclusion of the BNE and the NLS is rooted in our personal histories. Our native proficiency in Spanish, along with one of us being from Spain, facilitated our engagement with BNE’s digital initiatives, specifically the *Biblioteca Digital Hispánica* (BDH). Similarly, our previous residence in the United Kingdom sparked our interest in the NLS’s work. Initially drawn to the NLS through its Data Foundry initiative, our discussions quickly broadened to include the library’s comprehensive digital strategy, prompting us to incorporate the NLS into our study. These fortuitous circumstances enabled us to achieve an interesting European triangulation: a Nordic, an Iberian, and a British case.

4.3 Data Collection

To reach a better understanding of our cases, data were collected from two main sources: documents and semi-structured interviews.

Existing Documents

When speaking of documents, we refer to quite a heterogeneous category. In general, there is a multitude of options, varying from personal diaries (physical documents) to websites and blog posts (virtual documents). This shows that “document” can be understood in a much broader sense, as they can be either textual or visual material—if they “can be read”, they can be used as a source of data (Bryman, 2016). At the same time, the study of documents (and artefacts) is said to be a less intrusive method. They already exist—they were not generated for the sake of research—which means that the data have not been affected or influenced by researchers or even participants (Bryman, 2016; Wildemuth, 2017b).

This study used three types of documents as sources of data:

- Institutional documents: these are virtual documents found on the libraries’ websites or provided by participants. This category includes, but is not limited to, digital preservation policies, strategic plans and digitisation guides.

- Websites: these are considered virtual documents that can be studied using both quantitative and qualitative methods (Bryman, 2016). They were included as sources in this study because they contain relevant narratives connected to the digitisation projects and digital collections of the national libraries.
- Digital surrogates: whether regarded as virtual documents or digital artefacts, the way they are presented and published on the digital galleries is relevant to this study.

Interviews

The second data collection method used in this study was conducting semi-structured interviews. These types of interviews are like “purposeful conversations” that allow researchers to gain more insight about people’s perspective of a phenomenon or event (Luo & Wildemuth, 2017). At the same time, the structure of the conversation means that “researchers can keep an open mind” (Bryman, 2016, p. 10) as, even though there is a set of pre-determined questions, researchers can deviate from it, leaving room for unexpected topics to arise.

Incorporating this method into our study comprised several steps. First, we contacted potential participants, as it was important to get an insider’s perspective of the national libraries’ digitisation initiatives and projects. For this reason, we proceeded to identify relevant actors, selecting individuals whose roles were more strategic than hands-on, as this would allow us to get a more comprehensive view of the processes and rationale underlying them. Three participants, one per library, had this profile. A fourth participant—from the KB—was the head of a project and had a more hands-on involvement in the digitisation process. In all the cases, we cold-contacted them via email, using the addresses listed on the websites—a strategy that worked quite well.

Regarding the NLS, our first contact began at the early stages of this thesis. We initially wrote to the library’s digital scholarship librarian but, after receiving an out-of-office reply, were redirected to one of her colleagues—who, from now on, will be identified as P-NLS. This proved to be the beginning of a very fruitful conversation as, given her role at the NLS, she was able to provide a very detailed description and explanation of the digitisation strategies and processes at the library. The BNE participant, from now on identified as P-BNE, replied quite quickly and agreed on giving us an interview. Finally, the first participant from the KB, from now identified as P-KB1, was contacted via email but, after getting no response, we had to enlist the help of our supervisor. Once this was done, contact was established, and an interview arranged. The second KB participant, henceforth referred to as P-KB2, was contacted via the project’s official email address. We appreciate his willingness to assist us with this thesis.

Second, we developed an interview guide by preparing a list of 18 questions. These were formulated taking into consideration the research questions, relevant theory and our exploration of the national libraries’ websites and digital galleries. At the same time, we took this opportunity to address some points that were not covered by the documentation available or questions that

had emerged while studying the documents. The guide (see Appendix A) was written in English, and then translated into Spanish.

Third, we conducted the four interviews (1 NLS, 1 BNE, 2 KB). As the participants were in three different countries—Scotland, Spain, and Sweden—the interviews were conducted via video conference platforms (Zoom and Microsoft Teams) during March and April 2024. We were both present in the conversations, but only one of us guided each interview. The participants in Scotland and Sweden were interviewed in English, while the participant in Spain was interviewed in Spanish.

The final step was capture and pre-analysis. The audio from the interviews were recorded with the consent of the participants, and then transcribed using the proprietary software MAXQDA Plus 24. The full transcriptions were contrasted against the recording to check accuracy. Some words (e.g. fillers) and repetitions were deleted for the sake of the analysis but otherwise the texts are a literal transcript. Grammar and punctuation were not modified for two reasons: we did not want to interfere with the meaning of the message and editing the language would be too time-consuming. As it was, the content was clear and, if some interview snippets are used as examples, the fragments are edited when writing the report. Once this step was completed, the texts were ready for the analysis.

4.4 The Dataset

The National Library of Scotland (NLS)

Some of the documents used for analysis from the NLS are not publicly available and were made available to us by the participant. These will be marked as such in the list; the rest have been obtained from their website. Text file samples (.txt), METS.XML samples and ALTO XML samples were downloaded from NLS7, NLS8 and NLS9.

Institutional Documents

NLS1 – Digitisation request form for inclusion in the library’s mass digitisation programme (supplied by participant)

NLS 2 – Digitisation workflow for curators for paper-based materials (supplied by participant)

NLS3 – The Library’s approach to selection for digitisation⁶

NLS4 – Digital preservation policy (2023)⁷

NLS5 – Project digitise: How the National Library of Scotland is opening up its collections to all⁸

NLS6 – Open Data Publication Plan⁹

From the website and the digital gallery (samples)

⁶ <https://www.nls.uk/media-u4/1557983/2018-digitisation-selection-approach.pdf>

⁷ <https://www.nls.uk/media/uo4b1zjf/digital-preservation-policy-and-plan.pdf>

⁸ <https://www.nls.uk/media-u4/1624630/project-digitise-leaflet.pdf>

⁹ <https://nlsfoundry.s3.amazonaws.com/download/national-library-of-scotland-open-data-publication-plan.pdf>

- NLS7** – Intro to Early Manuscripts¹⁰
NLS8 – About Broad­sides Printed in Scotland 1650-1910¹¹
NLS9 – About Diary, Letter and Poems of Marjory Fleming¹²
NLS 10 – About Edinburgh Ladies’ Debating Society¹³
NLS 11 – “Decretum” of Gratian – Collection overview¹⁴
NLS 12 – “Decretum” of Gratian viewer¹⁵

From the website and the Data Foundry (samples)

- NLS13** – About the Data Foundry¹⁶
NLS14 – Data Foundry – Digital Resources¹⁷
NLS15 – Data Collections Standards¹⁸

Interview

NLS16 – Interview: conducted on March 12th, 2024, via Teams. Participant will be identified as P-NLS

The National Library of Spain (BNE)

The material gathered from the BNE consists of a publicly available documents from their website, and some forwarded to us by the participant.

Institutional Documents

- BNE1** – Digitisation process at the BNE: “Proceso de Digitalización en la Biblioteca Nacional de España: Biblioteca Digital Hispánica” (2024)¹⁹
BNE2 – Digitisation process at the BNE: “Proceso de Digitalización en la Biblioteca Nacional de España: Biblioteca Digital Hispánica” (2015). No longer available on the website.
BNE3 – Completed digitisation projects in 2023: “Líneas de digitalización completadas” (2024)²⁰

From the website (BDH) and the digital gallery (samples)

- BNE4** – Biblioteca Digital Hispánica homepage²¹
BNE5 – *Biblia de Ávila* [Vetus et Novum Testamentum cum praefationibus et argumentis Sancti Iheronymi et aliorum]: catalogue entry²²
BNE6 – *Biblia de Ávila* in viewer²³

¹⁰ <https://digital.nls.uk/early-manuscripts/>

¹¹ <https://data.nls.uk/data/digitised-collections/broadsides-printed-in-scotland/>

¹² <https://data.nls.uk/data/digitised-collections/marjory-fleming/>

¹³ <https://data.nls.uk/data/digitised-collections/edinburgh-ladies-debating-society/>

¹⁴ <https://manuscripts.nls.uk/repositories/2/resources/14983>

¹⁵ <https://digital.nls.uk/early-manuscripts/browse/archive/222543300#?c=0&m=0&s=0&cv=6&xywh=-767%2C-3651%2C11598%2C14080>

¹⁶ <https://data.nls.uk/about/>

¹⁷ <https://www.nls.uk/digital-resources/data-foundry/>

¹⁸ <https://data.nls.uk/about/standards/>

¹⁹ https://www.bne.es/sites/default/files/repositorio-archivos/proceso_digitalizacion_bne_2024.pdf

²⁰ https://www.bne.es/sites/default/files/repositorio-archivos/Lineas_digitalizacion_2023.pdf

²¹ <https://www.bne.es/en/catalogues/biblioteca-digital-hispanica>

²² <http://bdh.bne.es/bnearch/detalle/bdh0000014221>

²³ <http://bdh-rd.bne.es/viewer.vm?id=0000014221&page=1>

BNE7 – *Biblia de Ávila* download options²⁴

BNE8 – *Ethica ad Nicomachum*. Politica. Oeconomica [Texto impreso] / trad. Leonardo Aretino²⁵ catalogue entry

BNE9 – *Ethica ad Nicomachum* in viewer

From the website (Hemeroteca) and digital gallery

BNE10 – About Digital Periodical and Newspapers Library²⁶

BNE11 – *Diario de Granada* 1808²⁷

BNE12 – Sample of Digital gallery view of *Diario de Granada*²⁸

From the website and BNELab (samples)

BNE13 – BNELab Data²⁹

BNE11 – About BNELab³⁰

BNE15 – Linked data at the BNE³¹

BNE16 – *Diario de Granada* in datos.bne.es³²

BNE17 – Hispanic Digital Library datasets selection page³³

BNE18 – Hispanic Digital Library dataset selected³⁴

Interviews

BNE19 – Interview: conducted on March 15th, 2024, via Teams. Participant will be identified as P-BNE.

The National Library of Sweden (KB)

The materials we used to study the KB include, similarly to the other two national libraries, documents provided by the two interviewees, which are not publicly available, as well as a series of documents available in their website. The examples on Manuscripta and Gotlandslagen are representative. We looked at many more, however, we consider these to exemplify the type of documentation available in the KB.

The digital collections in this library are, as one of the participants puts it, in a more “dispersed ecosystem”. Most of the digitised printed material can be accessed through Libris and Regina, and digitised newspapers have their own dedicated section on the website. Manuscripts and personal archives, however, use different platforms. Personal archives are hosted in Arken, while Manuscripta has its own website, manuscripta.se.

²⁴ <http://bdh-rd.bne.es/viewer.vm?id=0000014221&page=1>

²⁵ <https://bdh.bne.es/bnearch/detalle/bdh0000176302>

²⁶ <https://www.bne.es/en/catalogues/digital-periodical-and-newspaper-library>

²⁷

<https://bdh.bne.es/bnearch/CompleteSearch.do?showYearItems=&field=todos&advanced=false&exact=on&textH=&completeText=&text=diario+de+granada&pageSize=1&pageSizeAbrv=30&pageNumber=3>

²⁸ <https://hemerotecadigital.bne.es/hd/viewer?oid=0004353004>

²⁹ <https://bnelab.bne.es/en/data/>

³⁰ <https://bnelab.bne.es/en/about-bnelab/>

³¹ <https://datos.bne.es/inicio.html>

³² <https://datos.bne.es/edicion/bise0000175978.html>

³³ <https://bnelab.bne.es/en/dataset/hispanic-digital-library/>

³⁴ <https://datos.gob.es/es/catalogo/ea0019768-biblioteca-digital-hispanica-bdh-set-completo>

Institutional Documents

KB1 – Preservation policy: *Policy för bevarande av Kungliga bibliotekets informationsbestånd* (supplied by P-KB1)

KB2 – Digitisation policy: *Policy för digitisering av Kungliga bibliotekets samlingar* (supplied by P-KB1)

KB3 – Quality requirements for digitization of older daily newspapers: *Kvalitetskrav för digitalisering av äldre dagstidningar* (supplied by P-KB1)

KB4 – The Latin and West Nordic handwriting projects *De latinska och västnordiska handskriftsprojekten* (supplied by P-KB1)

From the website (Arken) and galleries

KB5 – Welcome to Arken³⁵

KB6 – About Arken³⁶

KB7 – Det följande handlar helt och hållet om Nietzsche³⁷

KB8 – Dublin Core 1.1 XML example³⁸

KB9 – EAD 2002 XML example³⁹

From the website (Manuscripta) and digital gallery

KB23 – Manuscripta homeplage⁴⁰

KB24 – Manuscripta example 1⁴¹

KB25 – Manuscripta XML example 1⁴²

KB26 – Manuscripta IIIF Manifest example 1⁴³

KB27 – Manuscripta example 2⁴⁴

KB28 – Manuscripta XML example 2⁴⁵

KB29 – Manuscripta IIIF Manifest example 2⁴⁶

KB30 – About Libris⁴⁷

From the website and data.kb.se

KB10 – About Data.Kb⁴⁸

KB11 – Aftonbladet - Libris⁴⁹

KB12 – Aftonbladet – data.kb example⁵⁰

KB13 – Aftonbladet – manifest⁵¹

KB14 – Aftonbladet coding formats for downloading⁵²

³⁵ <https://arken.kb.se>

³⁶ <https://arken.kb.se/about>

³⁷ <https://arken.kb.se/SE-S-HS-L41-1-11>

³⁸ https://arken.kb.se/SE-S-HS-L41-1-11;dc?sf_format=xml

³⁹ https://arken.kb.se/SE-S-HS-L41-1-11;ead?sf_format=xml

⁴⁰ <https://www.manuscripta.se>

⁴¹ <https://www.manuscripta.se/ms/100088>

⁴² <https://www.manuscripta.se/ms/100088.xml>

⁴³ <https://www.manuscripta.se/iiif/100088/manifest.json>

⁴⁴ <https://www.manuscripta.se/ms/100209>

⁴⁵ <https://www.manuscripta.se/ms/100209.xml>

⁴⁶ <https://www.manuscripta.se/iiif/100209/manifest.json>

⁴⁷ https://libris.kb.se/help/about_libris_eng.jsp?language=en

⁴⁸ <https://data.kb.se/about>

⁴⁹ <https://libris.kb.se/dwpgqn5q03ft91j#it>

⁵⁰ <https://data.kb.se/dark-100476>

⁵¹ <https://data.kb.se/dark-100476/manifest>

⁵² <https://data.kb.se/dark-100476>

- KB15** – Aftonbladet METS.XML example⁵³
- KB16** – Aftonbladet ALTO.XML example⁵⁴
- KB17** – Aftonbladet image jp2 example⁵⁵
- KB18** – Gotlandslagen⁵⁶
- KB19** – Gotlandslagen coding formats for downloading⁵⁷
- KB20** – Gotlandslagen image example tif⁵⁸
- KB21** – Gotlandslagen image example jp2⁵⁹
- KB22** – Gotlandslagen METS.XML example⁶⁰

Interviews

KB31 – Interview 1: conducted on March 21st, 2024, via Zoom. Due to time issues, we were not able to go through all the questions during the interview. For this reason, the participant provided the answers to questions 12 to 19 from the interview guide (Appendix A) via email on April 3rd, 2024. We decided to consider these two interactions as one unit because the transcribed interview and the email had a similar structure (question-answer) that could be combined. The participant will be identified as P-KB1.

KB32 – Interview 2: conducted on March 13th, 2024, via Zoom. The participant will be identified as P-KB2.

4.5 Data Analysis

The data collected was analysed using qualitative content analysis (QCA). This method is used to describe the meaning of qualitative data by assigning categories to different parts of the material. Schreier (2012) lists three key characteristics of QCA: it is systematic, flexible and works by reducing data. It is systematic because researchers must follow a series of carefully planned steps, including building a coding frame that yields consistent results. It is flexible because it can be applied to different types of textual and visual material and the coding frame is adapted for each case. Finally, it reduces data because it focuses only on the relevant aspects.

This method was deemed adequate for this study for several reasons, the most important being that, by eliciting meaning from our data, significant themes, patterns and associations were uncovered. The result of this was used to create detailed descriptions of the processes in each library and establish comparisons. At the same time, while other qualitative methods, including thematic analysis, are also suitable for this task, QCA provided the structure and strategies to keep our research focused and consistent—it helped us build a clear plan to manage both the amount and diversity of the material and attain useful results.

The steps followed were those described in Schreier (2012), including:

⁵³ https://data.kb.se/dark-100476/bib4345612_18730630_0_147.mets.metadata

⁵⁴ https://data.kb.se/dark-100476/bib4345612_18730630_0_147_0001_alto.xml

⁵⁵ https://data.kb.se/dark-100476/bib4345612_18730630_0_147_0001.jp2

⁵⁶ <https://data.kb.se/dark-17653561>

⁵⁷ <https://data.kb.se/dark-17653561>

⁵⁸ <https://data.kb.se/dark-17653561/003983112%2C6900001%2Cw%2C34%2C10v.tif>

⁵⁹ <https://data.kb.se/dark-17653561/003983112%2C6900001%2Cw%2C111%2C49r.jp2>

⁶⁰ <https://data.kb.se/dark-17653561/mets.xml>

- Selecting the material
- Building the coding frame (Appendix B)
- Dividing the material into units of coding
- Testing the coding frame (pilot stage)
- Evaluating and modifying the coding frame
- Main analysis
- Interpretation and presentation of findings

Coding Frames

The steps above reveal the importance of the coding frame in QCA. Schreier (2012) explains that this tool works as a filter, as it provides a structure and separates the relevant from the irrelevant data. Also, to yield good results, a coding frame must meet the following requirements: *unidimensionality*, where one dimension captures only one aspect of the data; mutual exclusiveness, where, within a dimension, a unit of coding can be assigned to just one subcategory; exhaustiveness, where all relevant material is coded; and saturation, where no category is left unused. These four requirements informed the construction of our tools.

Existing Documents

A challenge of working with existing documents is that they were not created for the purpose of our study. For this reason, it was of vital importance to create a system that would allow us to identify and interpret those instances where the material was connected to our research problem. We created a simple coding frame that allowed us to identify instances that mentioned or referred to provenance and paradata.

Interviews

The amount of data collected from the interviews called for the construction of one highly complex coding frame (several dimensions, several levels). It was built using a mixed strategy Schreier (2012), where the main dimensions were concept-driven and the subcategories data-driven—they emerged during the coding of the material. Up to this point, all our research had been conducted remotely. However, given the complexity of this task, the coding frame was built by meeting in person, which allowed us to ensure that both reliability and validity criteria were being met.

4.5 Ethical Considerations

As seen above, part of our research design entailed interviewing actors involved in the digitisation programmes of three national libraries. To maintain the integrity of our study, our interaction with the participants was informed by a series of ethical principles (Bryman, 2016): taking part in the study should cause no harm to participants and they must give informed consent; their privacy must be protected; and, finally, there must be no deception—researchers should be honest and transparent when it comes to the objectives of the study and the participants' role in it. These principles were reinforced by Högskolan i Borås' code of conduct.

In practice, this meant that from the very beginning participants were informed about our identities—our names and status—and the reason for contacting

them. Once they agreed to being interviewed, consent was obtained so that we could record the audio of the interviews. Each participant was sent a form, whose template was provided by our university, stating the type of data we would be collecting, the purpose of doing so, and how it would be handled. Some participants requested a copy of our interview guide, which we provided, or did pre-interviews to familiarise themselves with the nature of our study.

This study entailed no collection of sensitive personal data. The data collected through the interviews was handled according to our university's requirements. The audio recordings will be deleted once the project is completed. When it comes to reporting our results, we decided not to disclose the participants' names or specific positions in the libraries. No sensitive issues were discussed, but we still preferred to protect their privacy and merely give an outline of their roles.

4.6 Final Remarks

Before presenting the results of the analysis, we would like to address two possible weaknesses in our research design: the use of case studies and QCA. One reported weakness of case studies is the fact that their results cannot be generalised (Bryman, 2016; Choemprayong & Wildemuth, 2017). However, our research problem and questions make clear, from the very beginning, that the aim of this study is not to produce generalisations in order to create a theory. The objective here is to contribute empirical material, furthering the description and understanding of digital provenance and paradata in real contexts.

One of the issues connected to QCA has to do with data redaction. When the material is subjected to this process, it reaches a level of abstraction that erases some specificities of the material—as Schreier (2012, p. 30) puts it, a “potential multiplicity of meaning” is sacrificed. However, this is one of the trade-offs of this method; by abstracting the material, researchers can identify and work with the most relevant aspects of the data. To avoid this risk, we have followed the advice of Schreier (2012), and built a coding frame that finds a balance between being too general—dimensions are so broad that there is almost no differentiation—or too specific—so much, indeed, that it is impossible to find common points in the data.

The last point that must be addressed has to do with issues of validity and reliability. Validity, in QCA, is concerned with capturing what researchers actually set out to capture (Schreier, 2012, p. 175). To make sure that the analysis is not invalid, dimensions and subcategories must be carefully constructed so that they “represent the concepts under study” (Schreier, 2012, p. 175). Reliability refers to consistency: “[t]he criterion of reliability requires that your data and your findings are free of error” (Schreier, 2012, p. 26). This means that coding must be consistent, and researchers must agree on their practices. Assessing reliability in QCA is possible: a “coefficient of agreement” can be calculated (Schreier, 2012, p. 170) or coders can reach an agreement on the codes assigned to each segment (Schreier, 2012, p.175). To fulfil the validity and reliability criteria, several measures were taken. The

steps delineated above were carefully followed, a coding manual was produced (see Appendix B), and researchers agreed on all the codified segments.

5 Results

This chapter presents the results of our analysis, which have been organised into two sections: 5.1 deals with the existing documents and 5.2 with the interviews. When reporting the results, we decided to adopt a highly descriptive approach. This will prepare the ground for Chapter 6, where we will use the theory and previous research to discuss our findings more critically and answer the research questions.

5.1 Documents

The investigation included a variety of document types that provide a broad view of the standards and strategies employed by each institution. For example, official policy documents reveal the theoretical frameworks guiding digitisation efforts, while metadata standards documentation illustrates the specific schemas adopted to ensure the thorough recording of digital artefacts. Digital preservation guidelines offer insights into each library's commitment to maintaining the longevity and accessibility of their collections.

In conducting our research into the documentation practices of digital collections across the three national libraries, we found it necessary to establish a pragmatic approach given the vast array of materials available. Due to the extensive volume of collections digitised and documented by these libraries, it was beyond the practical scope of this thesis to conduct an exhaustive review of every individual collection and documents within it.

To effectively manage this challenge, we selected representative examples from each library. These samples were chosen because they provide a clear illustration of the various documentation strategies employed—highlighting both exemplary practices and areas where gaps exist. This method allows us to present a comprehensive overview that reflects broader trends across the libraries' digital preservation and documentation efforts without delving into repetitive or overly granular details.

The decision to focus on specific collections rather than a comprehensive survey was driven by the desire to maintain a manageable scope and to ensure that our findings remained relevant and insightful to our central research questions. This selective approach is advantageous in that it reduces redundancy and focuses on the most illustrative examples, but it also limits the breadth of our coverage.

Acknowledging these constraints is important to understand that, while thorough, our analysis does not encompass every digital collection maintained by the libraries studied. This limitation is identified as a potential area for future research, where subsequent studies could explore additional collections or reassess the institutions' evolving practices as they adapt to new digital preservation challenges and technologies.

This section is divided into three subsections, each focusing on a different library. It is important to note that the length of these subsections varies significantly. The uneven length of the subsections is primarily due to the

availability and publication of data. For instance, the Biblioteca Nacional de España (BNE) has not published any METS, XML, or similar documents. Consequently, the corresponding subsection is shorter compared to others where more comprehensive data was accessible. This structural choice reflects the differing levels of documentation available from each source, and we aim to provide as much detail as possible within the constraints of the available data. By acknowledging this at the outset, we aim to maintain transparency and provide context for the varying lengths of the subsections.

5.1.1 The National Library of Scotland (NLS)

In an era where digital transformation is paramount, the National Library of Scotland (NLS) stands out for its meticulous approach to documenting its vast and varied digital collections. Central to the library's strategy are the METS⁶¹ and ALTO⁶² XML standards, which are not merely acronyms but the backbone of the library's efforts to preserve the integrity and accessibility of digital content. METS, which encapsulates bibliographic, administrative, and structural metadata, and ALTO, focused on the detailed layout and textual content, together ensure that every facet of a document's digital life is captured— from its physical state to its digital reincarnation.

Our access to both publicly available and internal documents, provided courtesy of an NLS interviewee, has allowed us to observe these practices in a depth not available to the general public. This privileged view sheds light on the nuanced strategies NLS employs in documenting the digitisation processes, enriching our understanding of the library's commitment to transparency and accessibility.

Delving into Provenance and Paradata

The narrative of any historical item is incomplete without a thorough understanding of its origins and lifecycle. At the NLS, while the provenance documentation robustly outlines the historical context of collections such as the Broadside and Marjory Fleming's manuscripts, it reveals less about the prior custody of these items before arriving at the library. This partial visibility into an item's past is like opening a historical novel mid-chapter—informative, yet not wholly satisfying.

Moreover, the paradata, or the data about the process of digitisation, exhibits similar gaps. For scholars, knowing that a manuscript was digitised is one thing; understanding how—the specific scanners used, the settings employed, the decisions made to handle fragile pages—can profoundly impact their trust in and use of the digital facsimile. Currently, such details are sparsely provided, leaving a narrative gap that could otherwise enrich an academic's understanding of the digital artefact's fidelity to the original.

⁶¹ METS: Metadata Encoding and Transmission Standard – an XML-based framework used for encoding and managing metadata of digital objects in digital library projects.

⁶² ALTO: Analysed Layout and Text Object – an XML schema designed to store OCR text and the layout information of a scanned document's pages.

The Open Data Plan and Digital Preservation Policy: Pioneering Transparency

The NLS's Open Data Plan is a bold commitment to transparency, aiming to make its treasures not only accessible but also adaptable by the wider community. This plan is part of a broader vision that aligns with Europe's leading libraries, offering a blueprint for how public institutions can contribute to a global pool of knowledge.

Parallel to this, the Digital Preservation Policy provides a detailed framework that underscores the library's proactive stance on digital longevity. The policy outlines sophisticated governance structures and strategic objectives aimed at ensuring that digital formats are preserved against the relentless tide of technological obsolescence. The inclusion of a Digital Preservation Content Register, though not fully public, is a testament to the library's structured approach to safeguarding its digital legacy.

Tales from Specific Collections

Each collection at the NLS tells a unique story, not just through its content but through its journey into the digital realm. The Edinburgh Ladies' Debating Society Collection, for instance, reveals a rich saga of women's intellectual contributions in 19th-century Scotland, with detailed metadata enhancing the texts' accessibility and searchability.

In contrast, the 'Decretum' of Gratian, a 13th-century manuscript, offers a narrative of historical scholarship with its roots in medieval Europe. The provenance information connects this manuscript to significant historical figures and locations, providing a vivid picture of its journey through time. However, the lack of detailed paradata for such a pivotal manuscript means that while the story of its past is rich, the story of its digital transformation remains untold.

Innovative Documentation in the Marjory Fleming Collection

A standout example of the National Library of Scotland's innovative approach to digital documentation is seen in the "Diary, letters and poems of Marjory Fleming" collection. This collection is particularly notable as it represents the first dataset within the library's Data Foundry project created using Handwritten Text Recognition (HTR)⁶³ software. This cutting-edge technology was employed to transcribe approximately 50 pages of Fleming's diary using the Transkribus⁶⁴ platform, which enabled the development of a tailored model of her unique handwriting.

⁶³ HTR: Handwritten Text Recognition – a technology that uses machine learning to transcribe and digitise handwritten text from historical documents or manuscripts.

⁶⁴ Transkribus - A comprehensive platform for transcribing, recognising, and searching historical documents, offering tools for Handwritten Text Recognition (HTR), Optical Character Recognition (OCR), and collaborative transcription.

This model was subsequently applied to the entire collection, facilitating a highly accurate digital transcription of historical texts that were previously accessible only in their original handwritten form. This process not only enhances the readability and accessibility of these valuable documents but also opens up new avenues for academic analysis and public engagement by converting handwritten texts into searchable, machine-readable formats.

Documentation Practices Across Platforms

It is also crucial to note that the documentation standards and accessibility of the Marjory Fleming collection within the Data Foundry differ significantly from those presented in the Digital Gallery. While the Data Foundry employs detailed METS and ALTO XML files, supplemented by extensive paradata about the digitisation process, such as the use of HTR technology, the digital gallery's presentation of similar collections may not always provide the same level of detailed documentation or technical insight. This distinction underscores the library's varied approach to digital archiving and presentation, depending on the platform and the specific aims of the collection's digital transformation.

This methodological diversity, while reflective of the NLS's adaptive strategies to documentation, also highlights areas where consistency in digital practice could be beneficial. For researchers and the public, having uniform access to detailed paradata across all platforms would enhance the understanding of the collection's fidelity and the technological intricacies involved in its digitisation process.

The "Diary, letters and poems of Marjory Fleming" collection serves as a prime example of how the NLS is leveraging advanced digital technologies to preserve and make accessible the textual heritage of Scotland. The use of HTR technology in this collection not only demonstrates the library's commitment to digital innovation but also sets a precedent for future digitisation projects. By integrating sophisticated artificial intelligence techniques with traditional archival practices, the NLS is ensuring that these cultural treasures are not only preserved but are also made more accessible and useful for future generations.

The findings from the NLS reflect a library that is aware of its role as a custodian of history and as a pioneer in digital library sciences. The robust metadata frameworks and the forward-looking policies speak to a library that is preparing its collections not just for current consumption but for future generations. Yet, the narrative of digital transformation at the NLS is not without its needs for improvement—details on the digitisation processes and fuller provenance could further enrich the academic and public understanding of its collections.

5.1.2 The National Library of Spain (BNE)

The Biblioteca Nacional de España (BNE) embodies the scholarly endeavour of transitioning its vast and historic collections into the digital domain. This process, documented through the BNE's digitisation protocols, reveals a commitment to marrying the provenance of historical texts with the technical

rigor of contemporary archival science. When referring to the digital collections of the BNE in our study, we specifically focus on those housed within the Biblioteca Digital Hispánica (BDH), the digital library platform of the BNE.

Cataloguing Provenance and Paradata

Within the BNE's bibliographic catalogue, the provenance of items such as Aristotle's manuscripts is catalogued with precision, offering insights into the editorial lineage and historical trajectory of these works. The catalogue entries serve as a bibliographic bedrock, yet they provide only the surface layer of the item's narrative. The details of prior ownership, the manuscripts' journey to the BNE, and the context of their acquisition, critical components of provenance, are not as readily apparent.

Contrastingly, the paradata—the metadata that details the digitisation process—is sparse within the catalogue but becomes more visible through the Biblioteca Digital Hispánica (BDH). The BDH offers a deeper dive into the technical aspects of the digitised items, with high-resolution imaging accompanied by colour charts to ensure the accuracy of digital renderings. However, specific details such as the digitisation equipment used, resolution details, and colour profile settings are not explicitly documented in the publicly accessible interfaces. This omission leaves a gap in the full understanding of the digital surrogate's creation, a critical element for researchers reliant on digital fidelity and colour authenticity.

Technological Narratives

While the BNE has embraced technology in its representation of collections, as seen in the user interface of the BDH, the translation of this technological narrative into a comprehensive documentation practice is incomplete. This is evidenced in the differing presentations between the catalogue and the digital gallery. The former remains a bastion of traditional library science, while the latter ventures into the realm of user interaction and digital display without full disclosure of the digitisation methodologies employed.

Bridging Documentation Gaps

The BNE's initiatives, such as the BNElab (BNE11; BNE13), suggest a strategic push towards openness and innovation. Datos.bne.es promotes bibliographic data through Linked Open Data, but it does not necessarily provide real-time reflections of the catalogue, nor does it detail the digitisation paradata. BNElab, while fostering digital reuse and creativity, remains distinct from Data Foundry in its scope, serving as a catalyst for digital experimentation rather than a comprehensive data repository.

In summary, the BNE's approach to digitisation and documentation presents a rich academic subject. It stands at the forefront of digital archival practice yet exhibits areas where transparency in provenance and paradata could be enhanced. Our observations, distilled from an array of representative examples, pinpoint where the BNE excels and where the scholarly community may benefit from more detailed documentation of the digitisation process.

5.1.3 The National Library of Sweden (KB)

The National Library of Sweden (KB) demonstrates a structured approach to the digitisation and documentation of its collections, underpinned by comprehensive policies that guide both preservation and access. These practices emphasize the maintenance of collections and ensure their accessibility through detailed documentation, adhering to both national standards and international best practices. These policies foster a framework of reliability, standardised procedures, collaboration, and legal compliance, ensuring the library's practices are effective and consistent across various platforms and projects.

Provenance and Paradata Documentation

The KB meticulously documents the historical significance and context of items within its collection. Each record is detailed, offering insights into the material aspects such as binding, dimensions, and the historical journey of the items. This detailed provenance documentation helps in preserving the cultural heritage and providing context to the library's patrons, enabling a deeper understanding of the collection's value and origins. However, while the descriptive metadata is rich, the documentation on the prior ownership and the chain of custody before the items' arrival at KB could be more comprehensive. This gap in the provenance can limit the understanding of an item's full historical context, which is important for certain types of scholarly research.

In terms of digitisation, KB's processes are well-documented, with clear adherence to established guidelines and standards such as METS and ALTO, ensuring high-quality digital outputs. The documentation includes detailed records of digitisation decisions and activities, crucial for internal quality assurance and external transparency. These processes are supported by technical metadata that often includes information about the digitisation equipment used, resolution, file formats, and colour profiles.

However, there is a noted need for more detailed public metadata concerning the technical aspects of these digitisation processes. The public documentation often lacks specific details about the settings and parameters used during the digitisation, which are essential for researchers relying on digital surrogates for detailed analysis. Additionally, the documentation could better outline the conditions and limitations on user access to these digital files, enhancing clarity and trust in the digital resources provided by the library.

Provenance and Paradata in the Manuscripta Project

A notable example of KB's documentation excellence is the Manuscripta project, which catalogues medieval and early modern manuscripts across Sweden. Each manuscript entry in this project is accompanied by detailed TEI (Text Encoding Initiative)⁶⁵ XML metadata that includes exhaustive descriptions of the physical and textual characteristics of the manuscripts. This paradata provides insights into the digitisation parameters, such as scanner

⁶⁵ Text Encoding Initiative: a standard for representing texts in digital form, using XML to describe the structure and features of texts. It is often employed in digital humanities projects for encoding manuscripts, literature, and historical documents.

types, resolution, and file formats used, offering a clear window into the technical aspects of how these historical documents are digitised.

Furthermore, the Manuscripta project includes comprehensive provenance information that traces the history and ownership of these manuscripts before their acquisition by the library. This level of detail enriches the scholarly value of the digital surrogates, making them not only valuable resources for research but also trustworthy reproductions that maintain historical integrity. However, we observed that the level of documentation varied depending on what institution held a particular manuscript, highlighting the challenges and negotiations involved, including differences in digitisation practices and quality standards (Dillen, forthcoming).

Enhanced Documentation Practices at KB

The National Library of Sweden (KB) employs a sophisticated approach to digitisation and documentation, leveraging comprehensive policies to guide the preservation of and access to its collections. These practices, supported by principles of reliability, standardised procedures, and collaboration, ensure effective and consistent management of the library's resources.

Platform-specific documentation practices

- Manuscripta.se: Utilises TEI XML to provide exhaustive descriptions and provenance details, significantly enriching the scholarly value of digital surrogates.
- Data.kb.se, digital data depository: Similar in concept to platforms like the NLS' Data Foundry, data.kb.se is an initiative by the National Library of Sweden to provide a wide array of the library's data in machine-readable formats. This platform is designed to facilitate extensive scholarly activities and technological applications, promoting open access and encouraging the creative use of the library's collections. It exemplifies KB's commitment to making its digital resources widely accessible and useful for a variety of users, paralleling international trends in digital library services. This platform facilitates the use of the library's digital resources for a wide range of scholarly activities and technological applications, promoting open access and encouraging the creative use of the library's collections.

Addressing Gaps and Enhancing Transparency

While the KB excels in many areas of digitisation and documentation, there are opportunities for improvement, particularly in making these processes more transparent to the public. For instance, the detailed technical metadata and digitisation standards observed internally could be made more accessible to users, enhancing the transparency of the digital surrogates. Furthermore, clarifying the conditions and limitations on user access to digital files could improve the usability and trust in the digital resources provided by the library.

Legal Compliance and Access Regulations

The KB is vigilant in managing its legal responsibilities, ensuring compliance with copyright regulations and data protection laws. Regular assessments are conducted to confirm adherence to these laws, underscoring the library's

commitment to ethical practices. However, the transparency around these processes, particularly the conditions under which users can access digital materials, could be improved to provide clearer guidance to users and researchers.

Summary

The National Library of Sweden exhibits a strategic and principled approach to the preservation and digitisation of its collections. The emphasis on standards, thorough documentation, and legal compliance positions KB as a leader in the field. Enhancements in detailing the technical aspects of digitisation processes and providing more explicit documentation regarding access conditions would significantly improve the utility and transparency of their digital collections for a global audience.

5.2 The Interviews: Insiders' Perspectives

We are aware that these conversations cannot provide an exhaustive view of the processes, as each individual experiences them from their point of view. These conversations were nevertheless illuminating as they revealed the libraries' dynamics and routines. The results have been organised around the four main dimensions of our coding frame: "Documentation Practices", "What Is Documented?", "Reasons for Documenting", and "Challenges". Each of these, in turn, will be structured according to the subcategories that emerged during the analysis (for details, see Appendix B). The relevance of these subcategories will vary from case to case.

5.2.1 The National Library of Scotland

Documentation Practices

Provenance or paradata are captured through both automatic and manual methods. The library has its own digitisation workflow management system built in-house that captures technical information (e.g., machine, operators, dates) automatically. Data regarding "digitisation-related decisions," if included, must be manually entered. The team also uses Microsoft SharePoint and Microsoft Teams to manage and work on documents. These tools allow them to collaborate and keep track of the projects, as they can share files and edit live documents.

According to the participant (from now on identified as P-NLS), the team relies heavily on spreadsheets, which can be generated either by exporting data from the system or manually. These files contain collection-level and item-level information and, in addition to describing the objects, they can help track the process. P-NLS emphasises that the team is dealing with many files and the data are dispersed—there is an unknown number of spreadsheets containing information about the digital collections, objects and processes.

According to P-NLS, the department does not have a digital or digitisation strategy per se. What happens, instead, is that the digitisation process is shaped and guided by other institutional policies and position statements. Decisions are made at different levels and these institutional documents allow both individuals and teams to make informed decisions. The digitisation team has

established well-defined standards and guidelines, which vary depending on the type of project or material being digitised. These standards may align with external recommendations, but are still internally defined, adapting to the context and answering to the specific needs of a project. All this information, as well as additional instructions on how to carry out certain parts of the project, are documented in text files (e.g., a Word document).

When it comes to sharing and accessing information about object creation or a collection, P-NLS says that she imagines three concentric circles—each representing a level of access. The third (external) tier is the public, who can access a digital collection once it is published on the library’s website. At the collection level, users are given a brief “narrative” of the collection—an explanation in plain text that is published on the website—as well as rights and re-use information. At item level there is some metadata, but not much, and users will not find information about the digitisation process. When asked about including digitisation guides, P-NLS says that sharing these with the public would not be helpful, as the processes are varied and are constantly evolving:

I guess we could [publish a digitisation guide], to be honest with you (...) I would not know how long it would be current. Or accurate. (...) We don’t have one digitisation guide. What we have is for every digitisation format we have the most recent guide we have produced. (...) Usually when I say send our guides out, I mark them a little bit or add comments and saying ‘This, this is here because this was written for a particular collaboration.’ And just to explain why we have this in. But usually, we don’t comply with this you know. (...) So sometimes you set parameters and they stay like that for a long time. And sometimes you change [them] every few months. Little bits. (...) But I also wouldn’t necessarily (...) find it helpful to other than researchers. Yeah. What we do have online is our approach to selection for digitisation, which is, which is outdated, by the way. Well, it’s not that it is outdated. (...) Timeline wise, it is outdated, but the approach is still the same at the moment.

The second (middle) tier comprises library staff who have access to documents and data. The NLS has a digital asset management system (DAMS) with a user-friendly interface that allows those with access to it—who do not necessarily belong to the digitisation department—to consult information about digital collections or objects. They can get information about the creation process and history of a collection or object, but they can also read notes and instructions left by other members (e.g., “do not use this image because it is still under copyright”). Additionally, if somebody has questions, the information in the DAMS can point them to the right person.

The last tier, or “inner circle”, refers to the digital production team, who are interested in the workflow management data. This information is on SharePoint⁶⁶ and other library staff can potentially access it. However, P-NLS says that she cannot think of any reason why someone external to the team

⁶⁶ SharePoint, developed by Microsoft, is a web-based collaboration and document management platform. It is designed to help organisations create, manage and share content and information seamlessly across teams and departments.

would be interested in doing this, as the information in the DAMS should be enough for them.

The library's digitisation approach and practices have changed a lot in the last 15 years. P-NLS mentions that they now record much more data than in the past:

So I can't speak for the other two digitisation teams⁶⁷ but for my team, for that programme team [the Data Foundry], we now record that information by default. (...) We don't ask ourselves (...) will a dataset be created from this? Um, we just record it, again, in spreadsheets.

This change was partly inspired by the creation of the Data Foundry, the library's data initiative that pushes for the publication of open and transparent datasets (for details, see [Ames, 2021](#)). Using a suite of tools has allowed them to succeed at this task, as now they can record information about the process (e.g. equipment, responsible team, physical location) that before 2018 was not captured.

They also generate more documentation for their systems and workflows, including security guidelines and environmental impact considerations. This change was driven, in part, by the growth of the team. More stakeholders, with different professional backgrounds, have been invited to participate in the digitisation process, creating a need to communicate ways of working.

What Is Documented?

During the analysis, three subcategories emerged. The first one, which we named "Steps and handling," showed that the team document the steps in the digitisation process. Information about the physical objects, such as change of location or the assessment of the condition, is also recorded.

The second subcategory shows that data about the agents involved are also documented. The digitisation chain is quite complex, requiring different people to be involved in—and responsible for—different parts of the project. Machines are also agents in this process, so data about equipment and related tools are also recorded.

The last subcategory, "Not documented," refers to times when steps or decisions are not captured during the digitisation process. P-NLS explained that sometimes some steps in the process are changed. For instance, professional expertise and experience might justify changing strategies as a different one might work best in a specific case. The decision-making process and the rationale behind a deviation is not documented. In the same vein, tiny day-to-day changes are not documented, as these can be too many, varied and frequent. Our impression is that they do not want to hinder, interrupt or slow down the digitisation process. Instead, they want to design processes and strategies that run as smoothly as possible. Documenting these changes or more spontaneous decisions—which are nevertheless well-informed and supported by the staff's experience—might affect this goal.

⁶⁷ Here, P-NLS refers to other branches of the libraries: one that digitises audiovisual material and another that deals with public requests. P-NLS' team oversees the library's planned digitization programme: "any larger scale programme for paper-based collections".

Reasons for Documenting

Provenance and paradata are captured and documented for several reasons. The first has to do with (internal) information sharing. The digitisation team has grown, and the creation process has become more complex. Recording information makes things run more smoothly as it is easier to keep track of the projects and processes—so even if a staff member falls ill, for instance, somebody else can take over their tasks. P-NLS even mentions that communication runs so well, and practices are so well-established, that the process has become a learned behaviour for operators.

This subcategory also emerged because the staff writes down instructions and guides. These documents are created, in part, with the aim of helping with training and onboarding. New employees, regardless of their role in the chain, have access to them and can learn the process; the information sharing improves workflow. This information also helps clear misunderstandings between departments. P-NLS explains that curators ask for physical material to be digitised by submitting a form. The issue was that they were not aware of how long it would usually take to do this, and questions would arise. For this reason, the digitisation department created a document that explains the steps in the process and how involved curators are in each of them. This made the process more transparent, clarified time spans and helped manage expectations.

Transparency is another reason for documenting provenance and paradata. In recent years, ethical questions have been posed, especially by the Data Foundry staff, whose datasets—in the spirit of “collections as data” (Padilla, 2017; Padilla & Higgins, 2014)—are based on the libraries’ digital collections. The datasets contain structured and transparent data to facilitate research and data re-use. For this reason, they turn to the department that digitised the material asking questions about the collections’ provenance and creation.

The subcategory “Preservation” also emerged during the analysis. However, the conversation did not expand very much on this topic.

Challenges

The first subcategory that emerged is “Technological challenges”, which, in turn, can be further subdivided into two. First, the NLS faces challenges related to the interaction between old and new technologies. Migrating information from an old system to a newer one, for instance, involves some risks and on occasion data have been lost during the migration process. Incorporating new software into the workflow can also be a source of anxiety as older systems can be rendered useless or incompatible. Finally, P-NLS says that change will take place when tools, technology or resources no longer help fulfil a need. However, knowing how much can—or should—be done when introducing new tools or technology can be challenging.

Another issue relates to technology affordances—what the available technology and tools allow us or prevent us from doing. The library’s in-house digitisation tool, for instance, can record image capture data (e.g. information

about the equipment and settings) but cannot record the decisions that informed the process. To enter these data, the team must rely on humans, which, as will be seen below, can lead to errors or information gaps. At the same time, there is no centralised system gathering all the documentation. Instead, information is scattered across different systems (e.g. databases, spreadsheets, etc.) and only people can make the connections to create a more complete picture of the process.

Another observation is that available technology can sometimes limit actions or hinder change. Even though digital provenance and paradata are collected and stored in different files and tools, these cannot be provided to the user because the visualisation tool (digital gallery) is not designed to do so. The infrastructure of the library's website prevents this from being possible. Additionally, sometimes equipment or software are not customisable, so even if the technology is state of the art, changes cannot be introduced to meet the specific needs and workflows of the library.

Another challenge is the existence of information gaps, which sometimes can be caused by issues with files and formats. P-NLS says that some information about the digitisation process was recorded—it exists somewhere—but it cannot be (easily) retrieved. There are documents containing provenance and paradata, but in such a large volume that they have become unmanageable. What complicates the matter even more is that some are paper files while others are digital files containing only machine-readable information. Team members can also be responsible for these gaps. P-NLS explains that if information is not instantly recorded, when members of the team leave (e.g., they retire), the information is lost, which is why data should be recorded as soon as possible. The library's retention schemes can also present some risks. On paper, after a certain period of time, files that have not been used—are not “live”—are deleted. However, this policy is not always followed, which could lead to some accumulation of unnecessary files or the accidental deletion of important information.

As a final point, we must mention that the digitisation team might also face institutional challenges. Being part of an institution means balancing priorities and some changes—even if they are important for the department—cannot be implemented because other things must be done first, i.e. there are priorities. At the same time, we have seen that different stakeholders are involved in the digitisation process, which can lead to there being conflicting views on what is best for the process. This means that different actors must learn to communicate their point and make compromises.

5.2.3 The National Library of Spain: *Biblioteca Digital Hispánica*

Documentation Practices

The department relies on both automatic and manual methods to capture data and generate documents. Information that is captured automatically can be stored as structured data and kept together with master files. These files are never modified and, even though it is not specified, it is safe to assume that they are stored in a digital asset manager system.

There is also less structured data about the context, processes and objects that are captured manually by actors in the digitisation chain. Documents containing this information are stored in the department's network drive. According to the participant (from now on identified as P-BNE), in this drive one can find "work procedures", such as documents containing the department's digitisation guidelines and standards as well as tracking documents. The latter are particularly helpful when working on projects that extend over a long period of time. In the same vein, the team also tries to document decisions agreed on in meetings, especially when these involve projects that have been undertaken by third parties. However, it cannot be guaranteed that this will always happen—but, if written down, they will be stored in the department's network drive. Finally, it is worth mentioning that P-BNE has also created her own set of documents that are not stored in the department's network drive. For each project, there is a file where she writes down decisions, problems, meeting notes, etc.

The reason for storing documents, except those that contain sensitive information, in the department's network drive is so that all the members of the department can share, access and work on them. The problem with using network drives is that information siloes are created, as other library departments—which have their own drives—cannot access these documents and vice versa. This hinders, or at least slows down, fruitful collaborations—for example, between the digitisation department and those working on the data lab initiative.

Many of the documents stored in the drive are also published on the library's website. P-BNE explains that the library is a public institution, and, by law, it must make all its processes transparent, including what takes place in the *Biblioteca Digital Hispánica* (BDH). By being available online, documents help users understand, to some extent, the context and decisions involved in the creation of a digital object (e.g., strategic plan, annual objectives) as well as the process (e.g., *Proceso de Digitalización*; BNE1). The digitisation guide on the website provides a less detailed account of the process. P-BNE says that this was a conscious decision, as the aim is to communicate the process to users in an understandable and effective way. She also points out that there are some documents that the team has deemed relevant only for internal use, which is why they have not been shared with the public. However, should a user need more information about the digitisation process, they could request these documents.

Regarding the changes that documentation practices have undergone throughout the years, P-BNE says that, compared to the past, the ways in which documentation is managed has worsened. She explains that the library used to have a document manager system that allowed different departments to keep all the documents organised according to international standards. It also addressed issues such as version control. This is no longer the case, and the library now uses network drives to store the documentation, even though this was discouraged in the past.

What Is Documented?

Two subcategories emerged during the analysis: “Steps and handling” and “Not Documented”. The department has created detailed guides on how different materials should be handled and digitised. This becomes particularly important when it comes to working with third parties. When an external group oversees a project, the department documents the decisions made about the project and creates documents listing requirements and specifications. In turn, third parties must deliver reports about the processes that took place and equipment used, thus ensuring that guidelines have been followed and standards met.

Regarding the subcategory “Not Documented”, the participant mentions that while the department is very meticulous when documenting external work, when it comes to projects undertaken by the department matters are different. There is usually less documentation produced and decisions that involve internal actors are not always documented. Our interpretation is that this might occur because, as BNE employees, they are already acquainted with BNE standards and workflows. Similarly, as staff, they are already accountable for their work and so do not need additional supervision.

Finally, there is also a difference between documentation at collection level and at item level. Information about items and their creation process are not provided individually but usually in sets or grouped according to their format or material. Strategies to transmit information (e.g. communicating typography size to researcher by placing a ruler next to the incunabula) are implemented across the entire collection.

Reasons for Documenting

The four subcategories that emerged during the analysis were “Information sharing”, “Accountability”, “Transparency”, and “Preservation”.

Regarding the first subcategory, paradata are captured and documented so that they can be shared with external groups. As mentioned above, when third parties are involved, there are documents that communicate the project’s objectives, specifications, requirements, deliverables, etc. P-BNE also says that some documents, more specifically the digitisation guide, are published on the website so that smaller libraries can base their own practices on what the BNE does. Sharing information about the digitisation process, digital collections, etc., is done in order to benefit the work of other institutions, not the users (as single individuals).

“Accountability” also relates to the department working with external groups. When third parties are involved, the library requires that they provide detailed information regarding quality control, image editing stages, etc. This information is delivered in reports which are stored in one of the library’s drives or in the external group’s cloud system, although the second option increases the risk of losing this information at some point.

Related to “Transparency”, BNE aims to publish as much documentation as possible as the library is a public institution and must comply with transparency laws. This includes publishing contracts and procedural information on the website. Thus, the intention here is not to enhance the transparency surrounding the digital object; this is a byproduct of being accountable to the state.

The final subcategory that emerged concerns the preservation of digital objects. Preserving digital information requires generating metadata, some of which can be regarded as paradata. A preservation system checks for corruption and creates a checksum to ensure long-term preservation. This applies specifically to digitised books, not general documentation.

Challenges

The BNE faces technological challenges, specifically related to technology affordances. During the first part of our analysis, we saw that not much provenance or paradata are provided when examining a digital object in the viewer—even though this type of data exists. When asked about this, the participant explained that the way documents are published depends on the structure of the website. They use metadata to make the resource findable and accessible, but they are restricted by what the visualisation tool allows them to show. The same happens with the architecture of the website: a document can be published only if the website is designed to or capable of supporting it.

Another issue related to technology concerns the software used to store and manage documents, which needs to be kept up to date, otherwise documentation best practice will be affected. The participant explains that years ago the library used Alfresco as a document manager, but the library stopped receiving updates to the point that it felt that employees were doing their work despite the software, not because of it. Eventually it was abandoned, and documents began to be stored in network drives. Using network drives is not ideal as it leads to siloes and documents being scattered across the institution. With no document manager, it is difficult to store documents in a systematic and organized manner or keep track of document versions.

The second subcategory that emerged was “Information gaps”. There can be inconsistencies when it comes to documenting decisions that involve internal versus external actors. When decisions need to be communicated to third parties, these are well-documented. However, when it comes to internal work, sometimes actions are performed without leaving a record of the rationale behind them. In other cases, the BNE might provide guidelines to third parties without storing a copy. At the same time, some things are left undocumented because staff members must prioritise other tasks. If they have time, staff members will try to document the process, but as time is a limited resource, only the most important parts are documented. To fill in the gaps, staff members rely on colleagues instead of documentation. P-BNE says that she also uses the library’s website and social media to recover information.

Poorly documented internal processes can lead to loss of knowledge or hinder information-sharing. Indeed, P-BNE recognises the advantages of documenting

internal processes as it helps navigate the library's ever-changing environment and dynamics. Lack of documentation also translates to loss of time and inefficient use of resources, especially when decision-making information is missing. P-BNE says that she has spent many hours deciphering the rationale behind certain processes or doing research on ideas that others already explored (and sometimes discarded).

The final subcategory relates to institutional challenges. Specifically, there appear to be decisions from upper management that affect the work of the department and, by extension, the documentation process. Upper levels might define tasks or objectives that demand the team's full time and attention, leading to internal initiatives being abandoned.

In the same vein, institutional policies that support the works of different departments need to be in place. Moreover, initiatives should be backed and driven by the institution as opposed to individuals, as they risk being abandoned when these leave the institution. P-BNE gives as an example the lack of a document management system and how this affects the quality of the documentation. Using Alfresco was a decision pushed by one person and, when she left the library, the system stopped being maintained and eventually was rendered obsolete. P-BNE says that the library needs a global system for project management, but there is no point in trying to implement this change from inside the department. This is an institutional and policy issue and must be driven by institutional directive.

5.2.3 The National Library of Sweden: *Digitala Kollektioner*

We interviewed two employees of KB: the participant from Interview 1 (from now on identified as P-KB1) provided a more overarching view of the library's digital strategy, while the participant from Interview 2 (from now on identified as P-KB2) shared his experience working on Manuscripta.

Interview 1

Documentation Practices

Mechanisms have been put in place to prevent the loss of information set to be captured from the beginning of the process. For instance, technical metadata captures the production process, and these data are kept in the "preservation package" which is ingested by the digital archive system. In the words of the participant: "No information should be lost if the process is run according to the standards and procedures decided during the planning stage". If no provenance or paradata around an object is found, this is not because of an accident or mistake. Instead, it means that from the beginning these data were not intended to be captured.

The subcategory "Guidelines and standards" became quite prominent in the analysis. The library has strict protocols and standardised workflows for the different stages in the digitisation process, which vary based on the project and type of material. These workflows are written in text files, so tasks can be carried out by different people consistently. When it comes to the technical

framework used for image capture and treatment, the library follows FADGI, and external groups collaborating with KB usually do the same. At the same time, for each project, the library creates quality assurance documents.

Having these standard routines and guidelines, as well as data about what has taken place in different projects, helps plan future projects. The library collaborates with external groups and researchers interested in conducting digitisation projects based on the library's collections. For this reason, there is a "planning group" consisting of different stakeholders, that assesses proposals and their feasibility. They can give advice on how to plan a project when applying for grants, as it is possible for them to estimate, for example, how much time a project should take or its cost.

KB has some projects that are more in line with what has been called critical digitisation (Dahlström, 2011). One of these projects is Manuscripta, which will be the focus of Interview 2, where we can see that when the source material is more unique, workflows and strategies can be tailored to the characteristics of the item. However, exceptions or deviations from the normal workflow are still documented and specified in the project's documents.

"Dissemination" is discussed in relation to the users of a digital object or collection. Both provenance and paradata are captured during the digitisation process and the resulting documents have an internal use. However, as we observed when studying the digital collections, end users see none or very few of these data. If someone external to the library wanted to learn more about the creation process of a digital surrogate, they would have to contact the library and request more information.

The library has experienced some changes in their practices. For instance, KB used to publish collections that were not based on their own physical collections, but not anymore. There are several reasons behind this decision, including questions of origin and ownership and legal issues. The library is today responsible for only its own collections, whose creation and publication it has overseen. The only exception relates to the KB's data lab, but the participant emphasises that, in that case, no images are used.

In the long term, the library plans to build a new website that can host all the library's digital collections and projects. When this happens, it will be possible to determine if or how much information about the process KB will publish and communicate to users. In the meantime, different projects will provide different amounts of information—Manuscripta, for instance, is much more transparent in this regard, especially when compared to some of the digitised books (PDF files) that can be accessed through Libris.

What Is Documented?

Three subcategories emerged during the analysis. The first one, "Project Complexity", refers simply to the fact that the processes are nuanced and may vary from one project to another, with documents detailing this complexity. This helps execute current projects and might also help plan future ones.

The second subcategory, “Steps and handling”, concerns the documentation of the steps in the creation process and the handling of the material. P-KB1 said that documenting project experiences as much as possible is a long-standing practice. For instance, projects must comply with quality assurance specifications, and ensuring that these are being met will become part of the contextual information. At the same time, the technical metadata includes information about the digitisation equipment, such as camera performance or quality tests, allowing problems to be traced back to specific cameras or OCR programs.

The last subcategory that emerged was “Unexpected problems and deviations”. Problems in the digitisation system are documented and can be traced back and corrected. Regarding deviations from standard practices, actors are allowed to modify processes and introduce changes in the routines, although everything must be communicated. P-KB1 says that changes to routines or procedures are reported in meetings, though it is not clear whether these are included in meeting minutes or only shared in conversation.

Reasons for Documenting

Concerning the reasons that support the practice of documenting both provenance and paradata, the analysis showed two clear subcategories: “Transparency” and “Accountability”.

Regarding the first subcategory, P-KB1 mentions that changes in the heritage sector demand more transparency and transparent objects. Indeed, linking physical objects to their digital representations is crucial in the KB workflows. At the same time, during the digitisation process information loss, or more specifically textual loss, can occur; the participant gives as an example performing image capture on a manuscript. There are established protocols for handling information loss issues. This absence of information is communicated to the user by presenting the missing sections of text at the end, a decision that strengthens the levels of confidence of those using digital surrogates.

Regarding “Accountability”, the conversation revolved around the existence of protocols and standards, as well as quality assurance documents. The fact that they exist does not guarantee that they have been followed, so it is important to leave a record of what took place during the digitisation process, thus proving that all the requirements were met. This is done, in the words of the participant, in order “[t]o document how these standards reflect in the methods and procedures used in the project”. Additionally, these guidelines inform in-house practices, but they are also shared with partners and third parties.

Challenges

The same three subcategories that emerged when analysing BNE and NLS were present in KB. For instance, the Manuscripta team devised an effective strategy to deal with textual loss. However, the tools used to generate the metadata cannot capture this process automatically and this must be done manually. Finding a balance between automated and manual processes is

challenging but necessary, as actions performed by humans can slow down the production flow. P-KB1 explains this well when he says:

A central issue is the balance between the actual production flow (efficiency and volume) and the effort needed to enter information relating to the documentation into the production system (or any administrative system that supports the production system). That is to say, the more automatized the documentation process, the more information can be added. If information has to be manually handled by the person operating the process, it will inevitably affect the production capacity.

When investigating the digital collections and reproductions, we saw that—aside from the Manuscripta project—there was almost no provision of digital provenance and paradata. When asked about this, P-KB1 explained that this information is not shared on the website because the infrastructure does not allow it. Special platforms must be built to make it possible to include more information about a project’s digitisation process. Indeed, the plan is to replace “the normal presentation mode of [the] digitised collections”, PDF documents retrieved through the Union and Libris catalogues, with a new tool that brings all the collections together:

We haven’t had any sort of interest in developing websites for our collections up until now, basically, when we’re looking at finding various different ways. So the website, the ecology of our library is quite dispersed at the moment. You have specific presentations for specific collections.

The subcategory “Information gaps” was not prevalent in our analysis. The only issue discussed was in terms of whether there is provenance information and paradata that end-users might consider useful that gets lost due to lack of capture or documentation. The participant says he believes that there are losses but cannot specify what would get lost, as relevance will depend on the user’s needs: “[It is] difficult to specify as their characteristics to a large extent is related to the interest of the user”.

The final subcategory that emerged was “Institutional challenges” and dealt mostly with two issues. The first relates to the fact that in this enterprise there are several multidisciplinary teams and departments involved. Different stakeholders, with different backgrounds, must work together towards the same goal. The analysis showed that this requires having good communication skills, balancing priorities and making compromises. This means that some processes or new technologies will simply not be introduced, as other library departments do not have the capacity or the willingness to do so. P-KB1 mentions, for instance, that some conversations with the IT department were challenging:

But there [is] also sort of kind of [a] struggle between the IT people that they don’t want five gigabytes of metadata, but they want to keep it as simple as possible. And from our side we want to keep it as big as possible. So that’s sort of, there’s some negotiation going on.

The second issue relates to institutional identity and expectations regarding the role of KB as a national library. Currently, the library only digitises its own physical collections, although this was not always the case. However, given

that this is a public institution, there is the question of whether it should act as a repository for smaller libraries. On the one hand, KB is a tax-funded institution, and it would make sense to serve other libraries by providing, for example, the infrastructure to host their digital collections. On the other hand, KB is responsible (or accountable for) the material hosted and published. This means, to some extent, the library must have some degree of control in—or knowledge of—the processes that take place during the digitisation of an item or collection. Provenance also enters the discussion, as there are legal questions and copyright issues that could complicate matters.

Interview 2: Manuscripta

Documentation Practices

This project has built “internal digitisation database[s]” using Microsoft Access, which are used to plan the digitisation of each manuscript. These databases contain information such as the manuscripts’ general condition, technical data about image capture and some decisions made in relation to this. Those who take part in the project—cataloguers, conservators, photographers—can access and enter data into them. P-KB2 also mentions that conservators have built a separate database where they record a more detailed account of the manuscripts’ conservation treatments. However, he has no access to it, nor has he ever seen it.

We asked P-KB2 whether he keeps personal notes (e.g. a diary) on Manuscripta. This does not seem to be a common practice: “occasionally I’ve written some sort of work diary,” he said, “but not systematically”. Instead, he keeps Excel spreadsheets:

Each manuscript is listed and then I fill in certain data. I mean, when was [a manuscript] taken to this waiting room (...) digitised or (...) brought back, resolution, the name of the photographer because that’s also recorded in this sort of digitisation database.

The subcategory “Guidelines and standards” emerged during the analysis but was not prominent. There are guidelines that describe different processes as well as possible solutions when problems arise. For instance, a strategy to deal with textual loss during image capture has been devised and, as of the time of the interview, conservators were in the process of writing it down: “sort of a workflow—how to handle this problem when it occurs”.

When it comes to “Dissemination”, internally, members of the project have access to the databases. Manuscripta also has a GitHub repository that is publicly available, which means that anyone external to the team can see and download the manuscripts’ descriptions (XML files) or explore how the project has been designed and published.

The subcategory “Past to Present” reveals that, for the participant, this project has been a process of learning and growth, where practices have become more standardised and data more consistent:

When I started with the Greek project, the first project we had, I wasn’t sort of... I didn’t use more standardised ways to report data then as I do now. I [have] become more aware of the importance of

[this], for example, when I create these Excel files that I try to use, data validation.

This evolution did not happen in a void. It is safe to assume that it has been shaped by years of experience and experimentation, as well as by conversations and discussions that take place in the cultural heritage field.

Concerning future plans, the project's current website does not provide information about the creation process of the digital surrogates—so far, it only lists the cataloguers. P-KB2, inspired by Whearty's *Digital codicology* (2023), would like to start including information such as the name of photographers and conservators or date of digitisation. This would make more transparent the intricacies of this type of work and acknowledge the contribution of different actors. Additionally, he is considering the idea of informing the user about the way in which the digital reproductions were put together—especially if there were some deviations in the process. He speaks of the cases where mobile phones have been used to perform image capture. The resulting images are added last in the sequence, despite the original order, and structural metadata can be used to establish the link to the pages they detail. P-KB2 says that users of the digital reproductions can probably see and understand the arrangement, but including an explanation of it could be considered.

Finally, there are some changes that would improve workflows and lead to better use of resources. Right now, Manuscripta exists separately from KB and has its own digital infrastructure, including an image server. This means that P-KB2 cannot delegate part of the work and issues such as image file formats and compatibility become his responsibility when, under other circumstances, they would not. The library plans to implement, in two or three years, a digital infrastructure that would bring together all the collections. About this, P-KB2 says:

I would be able to scale down the image server because it's a bit of a double process, because now I need to convert the images into the IIIF compatible format, which is a Pyramid Tiff or a Jpeg2000. It would be good if I could remove that part from Manuscripta because (...) the technical infrastructure [of this project] should be mostly focused on the XML. It's an XML database.

This change would allow him to start focusing all of his attention on the cataloguing of the manuscripts.

What Is Documented?

We identified three subcategories from the interview: “Steps and handling”, “Actors involved” and “Not documented”.

Regarding “Steps and handling”, everything that takes place during the digitisation process is recorded. For instance, it was explained that before image capture, manuscripts are assessed by conservators to determine whether the object is suitable for digitisation, and this information is entered into the database. This assessment informs decisions and determines if or how the physical object will be digitised. Photographers also have a say when it comes to image capture and tackling the issues that arise in each case. Also recorded in the database is data collected during image capture—for example, technical

data such as the number of images or resolution, or problems experienced during this stage.

When it comes to “Actors involved”, the digitisation process has experienced some changes. P-KB2 says that, in the past, the process “has been more anonymous”. However, it is now possible to track who has been involved in the digitisation process and who is responsible for the digitisation—for instance, the name of the photographers.

The last subcategory, “Not documented”, mostly concerns the TEI files and how to communicate information loss. P-KB2 says that “[he] think[s] if there is extra loss, then it should be mentioned”. It is not clear whether this has become a practice, but he mentions that when cataloguing a manuscript, it is possible to include, for example, “the text that is missing in the image” as cataloguers can go back to the physical manuscript and “capture it”.

Reasons for Documenting

There are two subcategories that emerged during the analysis: “Project complexity” and “Information sharing”. Since the foundation of Manuscripta, there has been an increase in the number of manuscripts the team intends to digitise, for instance, the two most recent projects include around 500 items. This has led to the introduction of changes: “the process before the project,” explains the participant, “was a bit more ad hoc. Now they [KB] have tried to sort of streamline it because of the sheer volume. I think this is the largest project they have ever had digitising this kind of material”.

“Information sharing”, in this case, refers to how information circulates both internally and externally. The manuscripts’ TEI files are available on manuscripta.se and GitHub. They can be shared with external parties, who can re-use them. P-KB2 comments: “the TEI community encourages the dissemination of TEI files because that is the bread and butter of the whole community—that you should be able to re-use the TEI files and see how others have done them”. The fact that the data are in a machine-readable format also facilitates research. Indeed, this part of the analysis reminded us, once again, of the collections as data imperative. The participant says: “When we created the schema for Manuscripta, we were also thinking a lot about quantitative [studies] (...) that it should be useful in quantitative studies, that the data can be standardised”.

Challenges

Similarly to the previous cases, three subcategories emerged during the analysis: “Technological challenges”, “Information gaps” and “Institutional challenges”.

“Technological challenges” mostly related to technological affordances. P-KB2 says that initially the Manuscripta team intended to use Uppsala University’s ALVIN as a platform. However, once the project began, they realised this would not be possible. Part of the problem related to “the recording of codicological units”, as working with manuscripts entailed

producing detailed descriptions of the material, including re-bound volumes composed of two or more manuscripts. ALVIN's metadata schema was insufficient for this task. Thus, Manuscripta had to be published separately, as this made it possible to use their own TEI adjusted to fit the data.

The subcategory "Information gaps" differs slightly from what we saw at the NLS or BNE. The discussion around loss or lack of provenance or paradata due to human involvement only arose when P-KB2 was asked directly about this. He said that he has been part of Manuscripta from the beginning, and cataloguers and conservators are usually quite stable staff. When it comes to photographers, there seems to be more rotation. New ones will need training if they have never worked with manuscripts. When they leave, they will take this knowledge with them, but this did not seem to be a cause of concern as new staff can be trained.

What was discussed, instead, was textual loss at item level, which can occur during the digitisation of a manuscript. The team, however, crafts solutions to prevent this or deal with it once it happens. The connection to "Information gaps" is not in relation to textual loss itself but with the strategies used to address it. Introducing new steps into the digitisation process will yield more and different data. They must make sure that there are mechanisms in place that will allow them to gather this new information, or it will be lost.

The last category in this dimension, "Institutional challenges", was discussed in relation to the differences between the library departments. Even if they are participating in the same project, their views and perspectives will be shaped by their fields or disciplines. When asked if, based on his experience, documentation practices have evolved through time, the participant speaks of levels of awareness. He says that he thinks that in some fields, like IT, documentation has always been a concern, while in others, with a more "humanistic background", this has come with time.

Issues between departments can also arise when trying to introduce changes. P-KB2 says: "the main issues have been trying to leave the old digitisation processes and guidelines and accept the fact that you need to think more of the quality". He then adds that increasing image resolution was "a bit of a conflict" as not everyone saw the need for doing this. He was pushing for this change, as it would increase the quality of the images and lower the chances of having to re-digitise a manuscript in a few years. Conservators agreed with his proposal, but the digitisation and IT departments were opposed as they were considering issues other than quality, for instance, image size and presentation. In the end, resolution was increased after some negotiation and alignment of views.

6 Discussion and Conclusions

6.1 Documentation practices

One of the questions that motivated this study relates to the types of documentation pertaining provenance and paradata that are generated in these national libraries' digitisation projects. The answer did not surprise us, in the sense that our observations coincide with what we read in the literature and previous research. This is by no means discouraging—on the contrary, it allows us to trace the direction that the NLS, the BNE and the KB are taking and understand the importance that contextual information might have in the creation and use of digital collections.

In this thesis, we adopted Huvila's (2022, p. 32) "middle-range perspective to contextual information," where creation context encompasses metadata, paradata, provenance and other elements—all of which overlap at some point. This allowed us to pay attention to different aspects of the digitisation process and identify a wide array of documents that contained traces of provenance and paradata. There is quite a variety of documents, and the data they contain display different degrees of structure, ranging from information represented by a metadata scheme to being written down in plain text.

The results showed that documentation is quite similar among the libraries. Some of this is very structured, such as metadata (e.g. PREMIS) and XML files, and aim to describe the digital reproductions. A closer inspection of these documents will reveal pieces of the creation process, e.g., technical information about the equipment, image capture and so on. Ways of representing information might differ as digitisation teams create schemes that adapt to the material being digitised and the context in which this takes place. However, the function of these files is similar, so we consider this type of documentation a common point in all three cases.

Similarly, NLS, BNE and KB create and use databases and spreadsheets which also contain provenance and paradata. Given the nature of these, the information they contain is very much structured. However, some comments from the participants pointed out some weaknesses or risks. P-KB2, for instance, says that Manuscripta began when he was less aware of the importance of generating structured data, which has caused some issues in the present. Apparently, some data cannot be easily exported, so he uses his own spreadsheets instead. Spreadsheets also present their own challenges as staff must ensure that these are kept up to date—failure to do so will lead to information gaps.

Beyond the structured data, we also found various types of unstructured data that pertain to provenance and paradata. These include documents that capture the decision-making process and the reasoning behind the creation of a digital collection or object. For instance, proposals, meeting notes, and other similar documents are generated and kept. These types of documents provide a narrative that captures the context in which decisions were made.

Additionally, there are documents that help maintain the quality of the work and the workflow. These documents include standards and guidelines which are essential for ensuring consistency and fluidity in the digitisation process. Within this category, we also identified less formal documents such as instructions and notes that explain how specific tasks were carried out. Despite their informal nature, these documents play a crucial role in maintaining the independence of the process from individual staff members.

We also observed more formal institutional documents aimed at ensuring accountability. These include contracts with third parties, which explicitly state the responsibilities and deliverables of each party involved. Reports from third parties back to the libraries are also part of this group, demonstrating compliance with agreed standards and procedures. These documents are vital for reconstructing the creation context of digital objects and ensuring consistency and transparency in the digitisation process.

This comprehensive documentation approach aligns with the operational flexibility and innovation highlighted by Filosa et al. (2023). Similar to what the authors argue, libraries like KB exhibit an agility that is vital in managing the evolving landscape of digital preservation. This balance of automated and manual processes optimises efficiency and output, showcasing how structured and unstructured documentation work together to ensure high standards of digital preservation. Operational flexibility is essential, blending manual oversight with automated processes to address the dynamic needs of digital library services. By integrating various types of provenance and paradata, libraries can achieve a more complete and adaptable framework for documenting and preserving digital collections. Mak's theory of archaeology, particularly her discussions on material bibliographical objects, provides a deeper insight into the significance of comprehensive documentation. She argues that material bibliographical objects—whether physical or digital—carry with them layers of history and context that are crucial for understanding their provenance and significance. By applying this archaeological perspective, we can appreciate that the documentation of digital collections is not just about capturing data but about preserving the intricate histories and decisions that shape these collections.

The documentation practices observed in these libraries underscore the significance of capturing both structured and unstructured data to maintain the integrity and context of digital collections. This approach reflects Mak's emphasis on the importance of recording the material and contextual details of bibliographical objects. By doing so, libraries not only safeguard the authenticity and reliability of their digital collections but also create a richer, more nuanced understanding of the objects they preserve. This alignment with Mak's theory highlights the role of documentation in making visible the often-invisible processes and decisions that underpin the creation and maintenance of digital collections.

Understanding the types of documentation generated is just the first step. Equally important is examining the reasons behind the creation of this documentation. This leads us to our next research question: what motivates

these libraries to document provenance and paradata in their digitisation projects? By exploring these motivations, we can better appreciate the value and impact of thorough documentation on the usability, reliability and authenticity of digital reproductions.

Documentation plays a crucial role in the preservation of digital collections, ensuring their long-term accessibility and alignment with best practices and international standards. This future-proofing approach minimises the risk of epistemic failure (McCarthy, 2007), and is fundamental to safeguarding and providing access to cultural heritage for future generations.

Provenance and paradata are also essential for designing workflows and conducting day-to-day activities within libraries. By maintaining detailed records, libraries can streamline their processes and improve efficiency. Adherence to guidelines and standards ensures consistency and quality across collections, thereby enhancing accountability. This is consistent with the arguments presented by Conway (2010) and Sköld et al. (2020), who emphasise the role of detailed documentation in supporting the sustainability and reliability of digital archives.

Furthermore, libraries must comply with regulatory requirements, adhering to legal and ethical standards, particularly when handling copyrighted materials. By doing so, they ensure that their practices are not only legally sound but also ethically responsible, thereby upholding the integrity of the collections and the institution. This aligns with the discussions by Dunning et al. (2017), who highlight the significance of regulatory compliance in the preservation of digital cultural heritage.

Moreover, the inclusion of provenance and technical metadata significantly boosts the transparency of digital reproductions, fostering trust among researchers and the public. These detailed records not only make the digitisation process more transparent but also contribute to the publishing of collections as data, making them more accessible for a wider audience and usable to the research community. Additionally, detailed documentation enhances the reproducibility and verification of digital content, aligning with the findings of Huvila (2022), who emphasises the importance of transparency in enhancing the credibility of digital archives.

Our findings from the digitisation practices of national libraries resonate with Essen's (2020) assertion, illustrating how provenance and paradata documentation serve as foundational elements in the preservation of cultural heritage within the digital realm. This documentation builds a body of knowledge about processes that not only benefits the quality of digital objects and collections but also fosters research and provides a model for others to follow. Reflecting on Mak's (2014) work on authority and authenticity, providing detailed information on the creation process enhances the credibility of the library and in turn its authority and the authenticity of its digital collections.

Documentation is not merely a record of past activities but an accumulation of documents and information that will be invaluable in the future. The initiatives undertaken by these national libraries in their digitisation efforts serve as a foundational framework. This framework will be beneficial not only for their ongoing projects but also as a reference for smaller libraries looking to develop their own digitisation practices. By setting high standards, clear workflows and demonstrating comprehensive documentation processes, these national libraries pave the way for others to follow, ensuring that best practices in digital preservation are widely adopted and consistently applied.

The importance of documenting often-invisible processes, as highlighted by Mak's theory, underscores the necessity of capturing both the technical steps and the decision-making context in digitisation. Each library faces challenges in managing the extensive data generated, ensuring consistency and maintaining high quality standards. For example, BNE's sparse public documentation of technical details highlights a gap in transparency, which can obscure the understanding of the digital artefact's creation and challenge its authenticity and usability. These challenges necessitate ongoing negotiations to balance detailed documentation with the practicalities of large-scale digitisation projects.

By critically evaluating the processes of creating, annotating and disseminating digital objects, we recognise the implications for the perceived authenticity and scholarly value of digitised materials. The varying levels of detail in paradata across our case studies underscore the need for transparent documentation to support the reliability and usability of digital collections. Applying the archaeological lens allows us to dissect how these institutional decisions reflect and reinforce specific power dynamics and cultural values within the library sector.

The methods for creating and using paradata are diverse, ranging from manual documentation to sophisticated digital tools, indicating its adaptability to various research environments and purposes. Different methods of recording documentation have been clearly expressed by our interviewees, highlighting the importance of paradata in attaining transparency. This transparency helps users understand the production processes, methods employed and resources used.

Despite differences in the amount of published information, the libraries we studied generally share a common pattern regarding the availability of provenance and paradata information on their websites. In most cases, there is very little public information or provenance data shared with the end user. As discussed, provenance and paradata are often retained internally, limiting public access. One notable exception is in smaller projects like Manuscripta, which differs from other projects at the KB, having been conceived as an XML database to be shared with researchers interested in manuscript studies. However, even in Manuscripta, the publicly available information on provenance is limited. While the provenance of the physical object is provided, there is little additional information in the XML files about the equipment used or the personnel involved in image capture. Users can see who catalogued the

item and view the physical description, but other details remain sparse. This discrepancy highlights a broader trend within mass digitisation projects in digital libraries, where the focus is often on production volume rather than the provision of comprehensive provenance and paradata.

Our analysis allowed us to draw some comparisons regarding the motivation behind the documentation practices in each library, though these are limited to the extent of our data. These exhibit variations across the three national libraries, influenced by differing cultural, institutional and technological contexts. For instance, in the case of the BNE, there is a strong institutional focus, with the digitisation department's work closely tied to the library's role within Spain. The BNE, as the head of the Spanish library system, aims to model its digitisation processes for smaller libraries. This institutional responsibility is evident in the documentation of outsourced processes and the requirement for third parties to provide detailed reports, reflecting the library's accountability obligations. The KB, in a similar fashion, also reflects on the library's role and responsibilities in Swedish society. However, the discussion around the generation of documentation put more emphasis on the integrity of the processes and potential for collaboration. Lastly, the NLS comes across as more user-focused, emphasising rapid and practical access to digitised materials. This focus on accessibility explains the minimal provision of information about internal processes on their websites, as the primary goal appears to be to make as much material available to users as quickly as possible. The day-to-day work at the NLS highlights transparency and documentation, yet this is geared more towards enhancing user experience rather than detailing the digitisation processes.

Each library's approach to documentation reveals different priorities: BNE's focus on institutional accountability, NLS's emphasis on user accessibility, and KB's interest in fostering collaboration. Such variations underscore the diverse strategies libraries employ to balance user needs, technological capabilities and institutional priorities. This variability not only reflects the libraries' unique operational contexts but also their differing priorities in terms of user engagement and resource allocation.

Additionally, the potential pitfalls of inadequate documentation are significant, as highlighted by Dahlström (2011). In critical digitisation projects, tailored solutions and manual labour may not always be well-documented, leading to crucial institutional knowledge becoming locked in the minds of individual contributors, making it less accessible for future use. This underscores the need for comprehensive and consistent documentation practices to ensure that valuable knowledge is retained and accessible for future reference and research.

6.2 Challenges and limitations

Our second set of research questions relate to the challenges that creators of digital collections at these libraries face, and how these impact the documentation of the creation process. The results of our study, presented in Chapter 5, showed that the problems that arise are connected to technology,

human agents and institutions. Thus, challenges can be grouped roughly into three areas. In practice, however, these are deeply connected.

To begin with, we will address some issues related to technology. We observed that the available systems and tools can determine—or limit—how much provenance and paradata are captured and recorded. If these systems lack the capability to capture certain types of data, parts of the process will not be recorded, leading to information loss. Similarly, if mechanisms to organise, maintain and access documentation are not established, the entire effort might be rendered ineffective, as library staff will not be able to use the collected information efficiently or at all, thus producing an incomplete picture of the creation process.

This issue is evident at the BNE, where the absence of a document management system affects the integrity of the documentation. P-BNE mentioned that changes in a document are not systematically tracked, making it unclear which version is the latest or which one should be used. At the NLS, a large volume of old paper files is in storage, with data that has not been—and probably will not be—used. Retrieving and migrating data from an analogue to a digital system is complicated (Burgess, 2016, p. 9), making it a time-consuming and inefficient task for the digitisation team.

To avoid the accumulation of useless data, the digitisation process should not be documented just for the sake of documentation. Instead, there should be a motivation or purpose underlying the practices, as having a clear understanding of this would help address “[t]he problem of capturing what” (Huvila, 2022, pp. 34–35). The literature and our case studies show that there is no one-size-fits-all solution, as each library and project has its own needs and demands. In other words, institutional and project objectives can dictate the provenance and paradata that should be captured and, with this information in hand, an adequate technical infrastructure can be designed and built. Although P-NLS says that data at her department are captured by default, the analysis shows that this is not a haphazard decision. The team is not only safeguarding the integrity of the digital collections, but also preparing the collections so that they can be used and re-used on other initiatives of the NLS. This is, at least, what the existence of the Data Foundry proves, as the initiative has published several collections as machine-readable datasets and will continue to do so.

This is by no means exclusive to the NLS—the BNE and the KB are also building their own data labs. We did not delve too much into it, but P-KB1 touched upon this when he mentioned the newspaper project, where the library, in collaboration with an external party, is digitising Sweden’s newspapers. The documents related to this project (KB10-KB22) demonstrated how much care has been put into planning and documenting the digitisation process. OCR has been performed in this project to make the data machine-readable and reusable. However, [Jarlbrink and Snickars \(2017\)](#) study this case and see that given the number of errors in the OCRred text and a lack of information on how it is processed (i.e. absence of provenance but especially paradata), the data are rendered less reliable and usable for research.

Another point has to do with the digital infrastructure of the libraries. More specifically, we refer to websites and visualisation tools, which pose some challenges that might affect, if not the completeness of documentation, the way it is used. In this case, the main issue is not that infrastructure hinders the capturing of provenance or paradata. Instead, based on our observations, the problem is that this might determine how digital objects or collections are presented and how much provenance and paradata are provided to external users.

One of our initial—and extremely naïve—assumptions was that libraries were not documenting much provenance and process information, which explained the lack of contextual information in the description of collections and reproductions. This, of course, proved to be incorrect. Upon examining the libraries' practices, we found that the type of data relevant to our thesis is indeed being captured and documented. This is an intentional and systematic practice that generates a series of documents containing traces of the origin, history and creation process of the digital object. When combined, these documents allow us to reconstruct the context in which the digitisation process took place.

However, we wondered why, if the information exists and sharing it might enhance the depth of description and levels of trust in a surrogate, it has not been shared with end users. Other libraries, such as the Bodleian Digital Library, have embraced the creation of contextual data as an opportunity. Burgess (2016, p. 82), speaking of this case, mentions that at the Bodleian Libraries there are “ongoing efforts to construct rich and informative contextual framework for the digital objects in our collections and their exposure to users and digital services on the web”. Provenance—and, may we add, paradata—plays an essential role in this.

Before we proceed with this point, we must mention that we are aware that this issue might still be more of a theoretical than practical discussion. Provision of provenance and paradata when publishing digital collections is not the norm at libraries, which is something that Dillen (forthcoming) discusses in his analysis of *Manuscripta*. He writes that “even in the case of scholarly editions, attesting to the way in which the digital reproductions that are used in the edition were produced is not common practice” (p. 9). In the case of this thesis, we are mostly dealing with mass digitisation projects in digital libraries, so it makes sense that, so far, the focus has been on production volume rather than the provision of contextual information.

What the analysis revealed was not a lack of interest or disregard for the documentation of the creation process. The answer was much simpler: the libraries' websites and visualisation tools cannot incorporate provenance or paradata into the presentation of the digital objects. Even though the participants see the benefits of making this change, for several reasons this will not be implemented any time soon. Notably, the attitudes of the participants towards this matter vary along a continuum. Influenced by Whearty's (2023) perspective, P-KB2, from *Manuscripta*, showed a growing interest in giving visibility to the creation process. For instance, he was thinking about starting to

list the photographers alongside cataloguers, thereby acknowledging their role in the creation of the manuscripts' digital reproductions. P-NLS, on the other hand, did not show an urgency in publishing information, in the form of guidelines or manuals, about the creation process on the website. This divergence likely stems from the differing scales and objectives of their digitisation projects. P-KB2 is involved in a project that focuses exclusively on manuscripts and is one of several projects at the KB—functioning rather independently from the rest. P-NLS, on the other hand, participates in a much larger operation—the mass digitisation of the library's physical collections.

There is one final observation related to infrastructure that we would like to address. This refers specifically to the KB and the “dispersed ecosystem” (P-KB1) of its digital collections. The way these are presented varies significantly and as a result, users can access collections in a very fragmented manner, with disconnected groups of digital objects scattered across different platforms. This dispersion made it challenging for us to understand the organisation and structure of the digital collections initially. Over time, we discovered that the collections are distributed across various platforms, such as Arken from Umeå University, *manuscripta.se*, and library catalogues like Libris and Regina, where digitised items can be accessed as downloadable PDFs. This inconsistency in data presentation reminds us of the importance of effective organisation, interpretation and accessibility of digitised collections (Huwe, 2023).

At the KB, although the practices within individual projects are well-documented and orderly, inconsistencies arise in the overall presentation and documentation across the collections as a network. This fragmented approach can obscure the understanding of the digital artefact's creation, challenging its authenticity and usability. The archaeological framework helps us observe these practices, highlighting the need for comprehensive and transparent documentation to support the reliability and scholarly value of digital collections. This raises questions about how these libraries, in particular the KB, will manage to consolidate all their digital collections into one coherent environment in the future.

Another type of challenge we observed relates to the lack or loss of information due to human actions (which correlates with the subcategory “Information gaps” in our analysis). Data can be captured by combining automated and manual processes (Lemieux & imProvenance Group, 2016), which is what takes place in our three cases. There were some issues that we observed when discussing the manual processes and how failure to perform certain actions might generate gaps—individuals might forget to capture planned activities, be compelled to prioritise other tasks, or perform actions without documenting them due to their familiarity with the work. The risk this entails is that once they leave the library, their undocumented knowledge leaves with them. Therefore, it is crucial to establish systematic documentation practices that do not rely on individual staff members. Automating processes as much as possible, shifting reliance from humans to machines, can mitigate these issues—which takes us back to our discussion of technology. We are not speaking of excluding human agents from the process, though, as it is essential

to acknowledge that the digitisation process, along with its comprehensive documentation, will always necessitate a blend of human effort and machine input.

Nevertheless, the analysis showed that the NLS, the BNE, and the KB, because of their carefully designed workflows and guidelines, are well-prepared to address these challenges. These measures help minimise the number of errors that could affect the quality of a digitisation project, including its documentation. This brings us to the question of when data should be captured. Lemieux and imProvenance Group (2016, pp. 20–21) position this along a spectrum: at one extreme, there is the decision to capture data "at the point of creation", while at the other, teams might choose to adopt a "forensic approach". Planning for the future is one aspect of the problem, but the three libraries must also deal with the consequences of not having documented provenance and contextual information in the past. In this regard, both the NLS and the BNE were seen to adopt a "forensic approach", retrieving information from different sources:

So, yeah, we have issues with legacy where we do not know, where we do not have reliable information on when was this [item] even digitised. And we might know the year in which it was proposed because we find these old forms and someone remembers, "Ah, that was the year I was pregnant". Or, you know, that kind of associations. But it doesn't mean that we know when they were captured. So, you know, sometimes you're happy when colleagues do not follow strict retention scheme guidelines and keep their files for longer than they should, because then you can trace information. (P-NLS)

There is a great lack of documentation here, many times this is not recorded anywhere. So, in the end it is what I was saying, that when there has been no transfer of knowledge, when I need to know why and how a previous work has been done (...) I have to go to the people who have been in the department longer or even if it has been an important project that has had greater dissemination, I have to go to the news on our own website or to our social network channels. Because sometimes, when a project is particularly important, maybe a video has been made about it and so on. (P-BNE; our translation)

While these techniques might prove effective, they are not very efficient because, as seen in the interview fragments, staff needs to invest time and energy in conducting detective work. At the same time, one could question how reliable the information is. Part of establishing facts and building timelines involves relying on biographical information and human memory, which is fragile, explaining why we have spent so much time discussing the importance of documentation.

The discussion on efficiency leads us to the final point we wish to address, which relates to institutional challenges and how institutional directives contribute to preserving the integrity of the documentation generated during the digitisation process. As we mentioned earlier, P-BNE highlighted that the lack of an institutional policy on the organisation and management of documents impacts the future use of information and slows down processes within the digitisation department. In contrast, this concern is not observed in our study of

the KB. There is a clear stance when it comes to handling documentation, as proved by one of the documents provided by this P-KB1: *Policy för digitisering av Kungliga bibliotekets samlingar* (KB2). This text lists the five principles guiding the library's digitisation work, states what will be documented (reasons for selection and prioritisation, decision-making processes, workflows), and it emphasises that all documentation is preserved for future use. Consequently, there is a well-established framework from the top down, outlining how to handle and navigate these processes effectively.

6.3 Final thoughts and future research

The theoretical framework based on Foucault's archaeological method, as adapted by Mak, provided us a structured way to uncover the layered processes involved in digitisation, revealing the complex decision-making and cultural practices that shape digital collections. This framework helped us to not only look at the technical aspects but also to understand the broader socio-technical systems in which these digitisation projects operate. This perspective allowed us to critically assess which aspects of the original materials are emphasised or obscured in the digital realm and how this shapes users' engagement with digitised collections.

However, the application of this theoretical framework also presented some challenges. The reliance on the theoretical lens sometimes made it difficult to balance theoretical depth and practical insights. While the archaeological approach provided a rich conceptual framework, it occasionally required substantial interpretive work to connect theoretical concepts with practical findings. Despite these challenges, the framework's emphasis on the materiality of digital objects and the processes of documentation significantly contributed to our understanding of digital provenance and paradata.

Perhaps most strikingly, our research reaffirms the enduring significance of the human labour in the digitisation process, re-evaluating the role and recognition of the human element. Amid rapid technological advancements, the expertise, decision-making capabilities and even personal notes of individuals are central to the curation and management of digital collections. This human-centric approach is important for navigating the complexities of digital library services and for designing training and collaborative workflows that leverage human insights alongside technological capabilities.

Integrating these understandings with the emerging view of digital collections as data, our thesis posits a paradigm shift where digital artefacts are recognised not only for their reproductive power but for the data they encapsulate. This perspective reframes digital collections from static repositories to dynamic datasets ripe for analysis and creative exploration, expanding the role of libraries in the digital age. Furthermore, our emphasis on the thorough documentation of provenance and paradata underlines their importance in ensuring the authenticity and reliability of digital collections. By presenting a meticulous account of the origins, handling and curatorial decisions of digital items, libraries can fortify the trust of their users and enhance the scholarly value of their collections.

The tapestry of digital library documentation is one of complexity and continual evolution, woven with the threads of technological advancement, policy development and human expertise. We hope that our thesis enriches the academic conversation by spotlighting the imperative for comprehensive documentation practices and advocating for a progressive perspective that harmonises the stewardship of digital collections with innovative approaches to data analysis and curation. Through this lens, our work has tried to humbly contribute to shaping a future where digital libraries not only preserve history but actively participate in the creation of new knowledge.

In hindsight, there are some considerations to be made regarding our research design. While having multiple case studies allowed us to draw interesting parallels between the different libraries, finding patterns as well as differences that relate to the purpose and identity of each institution, we believe that, given the limited time available, it would have been more fruitful to focus on one library instead. Each library would have been a suitable case study by itself, and we would have had the opportunity to analyse more in-depth the collections and interview more actors from a single institution.

Bearing this in mind, future research could expand on our findings by conducting single case studies of the entire digital collection maintained by one of the libraries. This would address the limitation of our approach and provide a broader understanding of each institution's practices. This more focused scope would enhance the depth and applicability of the insights gained, offering a more complete picture of digital preservation practices in these institutions.

As an alternative, future research could consider expanding the scope to include other types of libraries, such as public and academic libraries, to understand how documentation practices vary across a wider range of institutions. Incorporating quantitative methods, such as surveys, could complement qualitative findings and provide a broader understanding of documentation practices and challenges. Longitudinal studies could offer insights into how documentation practices evolve over time and their long-term impacts on the usability and reliability of digital collections.

Exploring the impact of emerging technologies, such as artificial intelligence, on documentation practices could present innovative solutions for enhancing digital provenance and paradata. Additionally, future studies could focus more on the user perspective, investigating how different documentation practices affect the usability and trust in digital collections among various user groups.

As we conclude our exploration of digital provenance and paradata documentation within the National Library of Scotland, the National Library of Spain, and the National Library of Sweden, we reflect on our initial analogy of a travel journal. Just as a travel journal records every step, sight and experience along a journey, our research has documented the intricate pathways through which digital artefacts are created, maintained and utilised. We trust that our findings have contributed to illuminate how these practices are not merely administrative tasks but are important for preserving the authenticity, integrity

and usability of digital collections. These practices ensure that the digital artefacts housed within our national libraries do not become mere replicas but continue to serve as trustworthy and reliable sources of knowledge, embodying the rich heritage they represent.

References

- Ames, S. (2021). Transparency, provenance and collections as data: The National Library of Scotland's Data Foundry. *LIBER Quarterly*, 31(1), 1–13. <https://doi.org/10.18352/lq.10371>
- Bearman, D. A., & Lytle, R. H. (1985). The power of the principle of provenance. *Archivaria*, 21, 14–27. <https://archivaria.ca/index.php/archivaria/article/view/11231>
- Björk, L. (2015). *How reproductive is a reproduction? Digital transmission of text-based documents* [Doctoral thesis, Högskolan i Borås]. <https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-881>
- Börjesson, L., Huvila, I., & Sköld, O. (2022, October 15). *Information needs on research data creation*. University of Borås. <https://doi.org/10.47989/irisic2208>
- Börjesson, L., Sköld, O., & Huvila, I. (2020). Paradata in documentation standards and recommendations for digital archaeological visualisations. *Digital Culture & Society*, 6(2), 191–220. <https://doi.org/10.14361/dcs-2020-0210>
- Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.
- Burgess, Lucie. C. (2016). Provenance in digital libraries: Sources, context, value and trust. In V. L. Lemieux (Ed.), *Building trust in information: Perspectives on the frontiers of provenance*. Springer International Publishing AG. <http://ebookcentral.proquest.com/lib/boras-ebooks/detail.action?docID=4643679>
- Cameron, S., Franks, P., & Hamidzadeh, B. (2023). Positioning paradata: A conceptual frame for AI processual documentation in archives and recordkeeping contexts. *Journal on Computing and Cultural Heritage*, 16(4), Article 75, 1-19. <https://doi.org/10.1145/3594728>
- Choemprayong, S., & Wildemuth, B. M. (2017). Case studies. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (2nd ed, pp. 51–59). Libraries Unlimited.
- Conway, P. (2010). Preservation in the age of Google: Digitization, digital preservation, and dilemmas. *The Library Quarterly: Information, Community, Policy*, 80(1), 61–79. <https://doi.org/10.1086/648463>
- Couper, M. P. (2017). *Birth and diffusion of the concept of paradata*. 16. https://jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf
- Dahlström, M. (2011). Critical editing and critical digitisation. In W. T. van Peursen, E. Thoutenhoofd, & A. van der Weel, *Text comparison and digital creativity* (Vol. 1, pp. 77–97). BRILL. <https://doi.org/10.1163/ej.9789004188655.i-328.29>
- Dahlström, M., & Hansson, J. (2019). Documentary provenance and digitized collections: Concepts and problems. *Proceedings from the Document Academy*, 6(1), Article 8. <https://doi.org/10.35492/docam/6/1/8>
- Dillen, W. (forthcoming). Paradata for digitization processes and digital scholarly editions. In I. Huvila, L. Andersson, & O. Sköld (Eds.), *Perspectives on paradata: Research and practice of documenting*

- process knowledge* (Knowledge Management and Organisational Learning). Springer.
- Dunning, A., Smaele, M. de, & Böhmer, J. (2017). Are the FAIR data principles fair? *International Journal of Digital Curation*, 12(2), Article 2. <https://doi.org/10.2218/ijdc.v12i2.567>
- Essen, J. P. (2020). Building historical knowledge byte by byte: Infrastructures and data management in modern scholarship. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories* (pp. 89–102). Helsinki University Press. <https://doi.org/10.2307/j.ctv1c9hpt8.10>
- Filosa, M., Gad, U., & Bodard, G. (2023). Description, translation and process: Making the implicit explicit in digital editions of ancient text-bearing objects. In G. Bodard & C. Palladino (Eds.), *Can't Touch This* (pp. 51–76). Ubiquity Press. <https://www.jstor.org/stable/jj.10067049.8>
- Foucault, M. (1972). *The archaeology of knowledge*. Harper & Row.
- Hart, T. R., & de Vries, D. (2017). Metadata provenance and vulnerability. *Information Technology & Libraries*, 36(4), 24–33. <https://doi.org/10.6017/ital.v36i4.10146>
- Hirtle, P. B. (2002). The impact of digitization on special collections in libraries. *Libraries & Culture*, 37(1), 42–52. <https://www.jstor.org/stable/25548976>
- Huvila, I. (2022). Improving the usefulness of research data with better paradata. *Open Information Science*, 6(1), 28–48. <https://doi.org/10.1515/opis-2022-0129>
- Huvila, I., & Ekman, S. (2024). Documentation of data making, processing and use facilitates future reuse of research data: The CAPTURE project. *Huminfra Conference*, 26–30. <https://doi.org/10.3384/ecp205004>
- Huvila, I., Greenberg, J., Sköld, O., Thomer, A., Trace, C., & Zhao, X. (2021). Documenting information processes and practices: Paradata, provenance metadata, life-cycles and pipelines. *Proceedings of the Association for Information Science and Technology*, 58(1), 604–609. <https://doi.org/10.1002/pra2.509>
- Huwe, T., K. (2023, June). *Digitization Starts the Process, Metadata Drives It—ProQuest*. <https://www.proquest.com/docview/2825236540?parentSessionId=65xpbn8sVakpKg3Zs2Z03OJzdGM0sZuFQL4TtcY%2BvMk%3D&pq-origsite=primo&accountid=9670&sourcetype=Trade%20Journals>
- Jarlbrink, J., & Snickars, P. (2017). Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive. *Journal of Documentation*, 73(6), 1228–1243. <https://doi.org/10.1108/JD-09-2016-0106>
- Kreuter, F. (2018). Paradata. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave Handbook of Survey Research* (pp. 529–535). Springer International Publishing. https://doi.org/10.1007/978-3-319-54395-6_61
- Lemieux, V. L., & imProvenance Group. (2016). Provenance: Past, present and future in interdisciplinary and multidisciplinary perspective. In V. L. Lemieux (Ed.), *Building trust in information: Perspectives on the frontiers of provenance*. Springer International Publishing AG. <http://ebookcentral.proquest.com/lib/boras-ebooks/detail.action?docID=4643679>

- Luo, L., & Wildemuth, B. M. (2017). Semistructured interviews. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (2nd edition). Libraries Unlimited.
- Mak, B. (2014). Archaeology of a digitization. *Journal of the Association for Information Science and Technology*, 65(8), 1515–1526.
<https://doi.org/10.1002/asi.23061>
- McCarthy, G. (2007). Finding a future for digital cultural heritage resources using contextual information frameworks. In F. Cameron & S. Kenderdine (Eds.), *Theorizing digital cultural heritage: A critical discourse* (pp. 246–260). MIT Press.
<http://ebookcentral.proquest.com/lib/boras-ebooks/detail.action?docID=3338737>
- Merriam-Webster. (2024, February 7). *PROVENANCE*. <https://www.merriam-webster.com/dictionary/provenance>
- OAIS Reference Model (ISO 14721)*. (n.d.). OAIS Reference Model (ISO 14721). Retrieved 12 February 2024, from <http://www.oais.info/>
- Padfield, J., Kontiza, K., Bikakis, A., & Vlachidis, A. (2019). Semantic representation and location provenance of cultural heritage information: The National Gallery Collection in London. *Heritage*, 2(1), Article 1.
<https://doi.org/10.3390/heritage2010042>
- Padilla, T. (2017, February 15). *On a collections as data imperative*.
<https://www.semanticscholar.org/paper/On-a-Collections-as-Data-Imperative-Padilla/ea3bd8c2111ab56cc656df5de4389b096a6d5ffc>
- Padilla, T. G., & Higgins, D. (2014). Library collections as humanities data: The Facet Effect. *Public Services Quarterly*, 10(4), 324–335.
<https://doi.org/10.1080/15228959.2014.963780>
- Scheltjens, W. (2023). Upcycling historical data collections. A paradigm for digital history? *Journal of Documentation*, 79(6), 1325–1345.
<https://doi.org/10.1108/JD-12-2022-0271>
- Schreier, M. (2012). *Qualitative content analysis in practice*. SAGE.
- Sköld, O., Börjesson, L., & Huvila, I. (2022, October 15). *Interrogating Paradata* [Text]. University of Borås.
<https://doi.org/10.47989/colis2206>
- Symonds, E., & May, C. (2009). Documenting Local Procedures: The Development of Standard Digitization Processes Through the Dear Comrade Project. *Journal of Library Metadata*, 9(3–4), 305–323.
<https://doi.org/10.1080/19386380903405207>
- Tognoli, N., & Guimarães, J. (2019). Provenance as a Knowledge Organization Principle. *KNOWLEDGE ORGANIZATION*, 46, 558–568.
<https://doi.org/10.5771/0943-7444-2019-7-558>
- VERBI Software. (2024). *MAXQDA Plus 24* [Computer software]. Berlin, Germany: VERBI Software. <http://www.maxqda.com>
- Warwick, C., Galina, I., Rimmer, J., Terras, M., Blandford, A., Gow, J., & Buchanan, G. (2009). Documentation and the users of digital resources in the humanities. *Journal of Documentation*, 65, 33–57.
<https://doi.org/10.1108/00220410910926112>
- Whearty, B. (2023). *Digital codicology: Medieval books and modern labor*. University Press.

- Wildemuth, B. M. (2017a). Descriptions of phenomena or settings. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (2nd edition). Libraries Unlimited.
- Wildemuth, B. M. (2017b). Existing documents and artifacts. In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science* (2nd edition). Libraries Unlimited.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), Article 1. <https://doi.org/10.1038/sdata.2016.18>

Appendices

A: Interview Guide

Specific questions regarding the person's role and work

1. Can you describe your current role and responsibilities at your institution?
2. What has been your role in the digitisation project(s) we are discussing today?

Documentation Practices

3. How many people are usually involved in the process of digitising a collection? What measures does the team take to ensure that all the participants are aligned, meeting the same standards and working towards the same goal?
4. Are those involved in each project required to document their work and rationale behind their decisions? If so, what and how? And how can other members of the team access this information?
5. When it comes to your specific role/work, do you keep either a formal or informal record of your actions or the decisions you have made regarding the digitisation project(s)? Could you elaborate and give some examples?
6. If or when parts of the project need to be outsourced, are third parties required to document their processes and work with the material? If so, how do they deliver that documentation to your institution?
7. If or when the project includes previously digitised material, what type of documentation regarding processes (paradata) and provenance is generally available? Do you incorporate this documentation into the current project? How?
8. How do you decide what information (paradata and provenance) should be included in the documentation of these projects? Does this vary depending on the type of project and/or initiative?
9. Could you elaborate on any significant trade-offs you have faced while deciding what information (paradata and provenance) to include in the documentation? How are these decisions navigated?
10. To what extent do you think your current metadata scheme allows you to include this information (paradata and provenance)?
11. Have you had to employ any supplementary strategies to document information (paradata and provenance) that cannot be adequately captured by the existing metadata formats? Could you describe them?
12. Who can access the documentation (related to paradata and provenance) generated during the digitisation process? And through what mechanisms?
13. What factors influence the decision to make documentation publicly available, available upon request, or restricted to internal use only?
14. When it comes to publishing the digital collection, how do you decide what (and how much) information (paradata and provenance) should be available for the users?

Challenges, Limitations, and Reflections

15. What were some of the main challenges or limitations you encountered in documenting the digitisation process? This, as an individual but also as a team.
16. How much information regarding processes and provenance do you think is lost during the project? Why do you think this happens?
17. In what ways have limitations in technology or resources impacted your ability to document provenance and paradata comprehensively?
18. Are there any resources or tools you wished you had access to that would have made the documentation process easier or more comprehensive?
19. Reflecting on your experience, how have documentation practices in digitisation projects evolved during your career?

Appendix B: Coding Frame for the Interviews

The table below contains the codes (categories and subcategories) used to analyse the interviews. The four main dimensions were created based on the theory and research questions. The subcategories, on the other hand, emerged from the data.

Each code is accompanied by a brief description, and the table was used as a coding manual. This helped us maintain consistency in our analysis.

D1 Documentation Practices	This dimension refers to the documentation practices that take place or have been established in the institution and digitisation team. In other words, ways of doing things and rationale.
<i>Changes in practices</i>	This refers to the changes the digitisation process (workflows, practices, behaviours) has undergone or will undergo.
Past to present	The changes have already occurred and been implemented.
Future plans	Changes have been thought but not yet implemented; desired or ideal scenarios.
<i>Guidelines and standards</i>	This refers to the guidelines and standards that inform the digitisation process and support the decisions made by the team.
Internal	This refers to the guidelines and standards written or created by the institution or team (even if they include some external sources). A more "bespoke" process based on the demands and needs created by the context.
External	This refers to standards and guidelines that were not created by the institution or team, e.g. international organizations or third parties (businesses).
<i>Dissemination</i>	This refers to the way (and how much) process and provenance information is made available and shared.
Externally	How information is shared, disseminated, communicated to library users.
Internally	How information is shared, disseminated, communicated within the institution (between staff members, teams, or departments).

<i>Recording data</i>	This refers to the way data regarding the process of creation and provenance of the digital object.
Automatically	The data is recorded automatically by machines and systems; humans do not interfere in the process of capturing and storing it.
Manually	The data is recorded manually; humans perform data entry tasks.
D2 What is Documented?	This dimension refers to what information regarding the digitisation process or provenance of the digital object is documented.
<i>Unexpected problems and/or deviations</i>	When something unexpected happens and part of the process differs from what normally should happen
<i>Steps and/or handling</i>	To leave a record of the steps that took place when creating the digital object and how the material was handled
<i>Actors involved</i>	Who took part or was responsible in (that part of) the process.
<i>Not documented</i>	Instances where it has been decided not to document certain actions.
D3 Reasons for Documenting	This dimension refers to the reasons that have led the institution, team and/or individual to document information regarding the process and/or provenance of the digital object.
<i>Accountability</i>	The process is documented to keep track of standards and good practices, quality control, especially when it comes to third parties involved.
<i>Project complexity</i>	An increase in the complexity of the digitisation project (initiative; chain) has led to the process being more documented.
<i>Preservation</i>	Process and provenance are documented for preservation purposes.
<i>Transparency</i>	Process and provenance are documented to make the digital object more transparent.

<i>Information sharing</i>	The process is documented to communicate information within the team (or areas) and improve workflow (smooth).
D4 Challenges	This dimension refers to the challenges that appear when documenting the digitisation process and/or provenance of the digital object.
<i>Technological</i>	This refers to the instances in which the documentation of the digitisation process has been hindered or affected due to issues related to technology.
Old vs. New technology	There is a struggle or clash between old and new technologies.
Affordances	Some things haven't occurred yet because the current technology cannot do that yet; there's no capability.
<i>Institutional</i>	This refers to challenges that arise because of institutional issues (top-down decision-making; interdepartmental communication; support; costs and funding, etc.).
<i>Information gaps</i>	This refers to those instances in documentation of the digitisation process and/or provenance of the digitisation process is incomplete, or doesn't exist, because information was lost or never captured in the first place.