Upcycling historical data collections. A paradigm for digital history?

Journal:	*Journal of Documentation*
Manuscript ID	JD-12-2022-0271.R1
Manuscript Type:	Article
Keywords:	Digital history, Reference works, Legacy collections, Retrodigitization, FAIR principles, Data Collection, Data rescue, Information creation, Paratext, Paradata

Upcycling historical data collections. A paradigm for digital history?

Structured Abstract

Purpose

Upcycling is conceptualised as a digital historical research practice aimed at increasing the scientific value of ~~texts~~ historical data collections produced in print or in electronic form between the eighteenth and the late twentieth centuries. ~~Applied to historical data collections, t~~The concept of upcycling facilitates data rescue and reuse as well as the study of information creation processes deployed by previous generations of researchers.

Design/methodology/approach

Based on a selection of two historical reference works and two legacy collections, an upcycling workflow consisting of three parts (input, processing and documentation, and output) is developed. The workflow facilitates the study of historical information creation processes based on paradata analysis and targets the cognitive processes that precede and accompany the creation of historical data collections.

Findings

The proposed upcycling workflow furthers our understanding of computational methods and their role in historical research. Through its focus on the information creation processes that precede and accompany historical research, the upcycling workflow contributes to historical data criticism and digital hermeneutics.

Originality/value

Many historical data collections produced between the eighteenth and the late twentieth century do not comply with the principles of FAIR data. The paper argues that ignoring the work of previous generations of researchers is not an option, because it would make current research practices more vulnerable, and would result in losing access to the experiences and knowledge accumulated by previous generations of scientists. The proposed upcycling workflow takes historical data collections seriously and makes them available for future generations of researchers.

1. Introduction

Since its inception in the early 2000s, upcycling has become a buzzword among entrepreneurs and environmental activists, eager to underline the forward-looking, sustainable nature of their product or initiative (McDonough & Braungart, 2013). Upcycling depicts the process of taking something old, outworn and seemingly worthless to make something qualitatively new, fashionable and of higher value (Zimring, 2016, p.45). In recent years, scientists have discovered the concept as a means to describe methods that move beyond data rescue and curation towards data reuse and integration (Vearncomble *et al.*, 2017; Avesani *et al.*, 2019). This paper argues that, due to its focus on scientific value gains, upcycling has a significant role to play in the historical sciences as a means to make historical data collections interoperable and reusable, on the one hand, and to examine the methods of previous generations of researchers, on the other hand. Dealing with historical data collections, upcycling also calls for a more general reflection on the achievements and remaining challenges of mass digitization.

The paper conceptualizes upcycling as a digital historical research practice aimed at increasing the scientific value of texts that have been produced in print or in electronic form between the eighteenth and the late twentieth centuries. Following Don McKenzie's seminal work, 'text' is understood as "(...) verbal, visual, oral, and numeric data, in the form of maps, prints, and music, of archives of recorded sound, of films, videos, and any computer-stored information, everything in fact from epigraphy to the latest forms of discography" (McKenzie 1999, p.13). McKenzie's broad definition of 'text' highlights the significance of non-book texts as one of many kinds of recorded forms (McKenzie 1999, p.12). It indicates that, beyond the examples of this paper, upcycling may be applied in various other research contexts as well, such as oral history (Thieberger 2018; McCartney 2018), film studies (Cherchi Usai 2010; Ricci 2008), or even 3D modelling (De Kramer 2022). ~~Within this broad definition,~~ This paper ~~upcycling~~ targets historical data collections (short: HDCs) as a particular kind of text. HDCs compile, structure and interpret data, typically (but not necessarily) from multiple sources. Many HDCs were produced before the invention of the computer; many more were produced before the era of laptops, digital cameras, stable internet connections and mass digitization. HDCs comprise *reference works*, such as dictionaries, lexica and encyclopaedia, and *legacy collections* that document the results of historical research projects focused on collecting data from primary sources.

2. Historical data collections

2.1. General characteristics

HDCs do not contain *raw data* like primary sources in archives, but *processed data* taken from these primary sources. A significant feature of both raw and processed data is that they are

meaningless when observed in isolation, i.e. as atomic, individual data items. If, however, data items are interpreted in relation to their surrounding data items, they become information. This requires structuring the data~~, for example, horizontally and vertically~~. We call this process information creation. Research based on information, or information analysis, generates knowledge. In the context of HDCs, an inextricable link between *raw* and *processed* data remains, even if it is not explicitly documented and may have become invisible in the course of data collection. ~~HDCs were produced at a time when the mass digitization of archival collections had yet to begin; in fact, many HDCs were produced even before the invention of the computer.~~ As a result, primary sources often exist independently from the HDCs in which they were used. Today, the situation is different. On the semantic web, Linked Historical Data express this inextricable link and may thus be understood as the union of primary and secondary sources (Meroño-Peñuela & Hoekstra, 2014, p.286).

2.2. Reference works

Reference works that aimed at collecting and distributing (historical) data and knowledge about a variety of topics became increasingly popular since the eighteenth century. Their proliferation intertwined with advances in book production and trade (Baten & Van Zanden, 2008; Buringh & Van Zanden, 2009; Bellingradt & Salman, 2017), a growing interest in the collection and diffusion of 'useful knowledge', and the emergence of the 'first knowledge economy' in Western Europe in the eighteenth century (Mokyr, 2002; Berg, 2007; Jacob, 2014). Few of these reference works are available as digital editions, like Johann Georg Krünitz' *Ökonomische Enzyklopädie* (Krünitz, 1773-1858).[1] A much larger number, for example, reference works in the realm of commerce and trade are merely available in print or as scans in digital library collections. Digital editions of merchant manuals, handbooks on weights, measures and currencies, and commercial practices, or of encyclopaedia for the handicrafts, industry and commerce are not available.

Sometimes, automatically generated OCR was produced during the scanning process. This offers some immediate new possibilities for working with historical reference works, but is insufficient for more encompassing analyses of their contents (***). There are two important reasons. First, as a subgroup of digital library collections for the humanities, historical reference works often use different character sets and have a complex page structure including tables, in-text columns and significant typographical features, which the commonly used (commercial) OCR pipelines do not handle particularly well (Cordell, 2017; Chiron *et al.*, 2017). With the broader category of humanities library collections, many ~~historical data collections~~HDCs share a 'circuitous path to digitization', which includes~~d~~ the production of

[1] http://www.kruenitz1.uni-trier.de/

scans from microfilms and has a negative impact on the OCR accuracy (Christy *et al.*, 2017). Second, a particular feature of historical reference works is that they hide a significant amount of implicit knowledge, which, due to organisation of the materials, dispersed across the entire work and is therefore almost impossible to retrieve using standard searches.

2.3. Legacy collections

The term legacy collections originates from archaeology, but is also used in other domains such as cultural heritage (Heitman *et al.*, 2017), library and archival collection management (Jaillant, 2022; Shiue *et al.*, 2021) and linguistics (Dobrin & Schwartz, 2021). Legacy, or 'orphaned', collections consist of "(...) older materials that do not meet modern 'best practice' curation standards and require considerable resources to be preserved for future research." (MacFarland & Vokes, 2016, p.161). As a rule, legacy collections were produced at a time when there was little concern for, or knowledge about, making data available to others (Vearncombe *et al.*, 2017; Clarke & Shiue, 2020). This is true for most legacy collections created before the year 2000 (cf. Kelly *et al.*, 2022)[2] and for much of the data and documentation collected by researchers in the course of their careers (Shiue *et al.*, 2021, p.308). In line with other disciplines dealing with legacy collections, many HDCs have "(...) lost curatorial support or w[ere] abandoned prior to transfer to a curatorial facility" after research results were published (MacFarland & Vokes, 2016, p.162), while the collections themselves remain "(…) as evidence of historical research and as potentially reusable data to support new scientific inquiries" (Shiue *et al.*, 2021, p.308). Insofar as HDCs are concerned, the term 'historical' has a double meaning, indicating both research that has been conducted in the past and is now historical and research about the past itself. HDCs are both. They contain evidence of the past, collected in the past, using past methodologies and techniques.

2.4. Where HDCs are kept

Many printed reference works and legacy collections are kept at memory institutions (galleries, libraries, archives and museums, short: GLAM), with many more hidden in personal libraries, index card systems, or computers. Thanks to large-scale efforts such as the digitization of sixteenth-, seventeenth- and eighteenth-century printed works in German-speaking countries (Sommer, 2010; Herr, 2012),[3] the Early English Books Online (EEBO) project,[4] or the Internet Archive,[5] a large and growing number of HDCs has resurfaced. They are available online as

[2] Today, most researchers collect their data digitally, which leads Kelly *et al.* (2022) to conclude that the issue with legacy collections is a finite problem. This, however, presupposes that researchers today apply accepted standards for data collection, which is not always the case.
[3] http://www.vd16.de/, http://www.vd17.de/, http://www.vd18.de/.
[4] https://proquest.libguides.com/eebopqp
[5] https://archive.org/

scans of book pages; sometimes complemented with automatically produced OCR. It is hard to estimate how many reference works would fall into this category, but even a quick look at some bibliographies and catalogues about a specialist topic, such as commercial dictionaries and merchant manuals printed before 1850, shows that their number easily runs in the thousands (see e.g. Besomi, 2012; Jeannin *et al.*, 1991-2001). Other HDCs have been deposited as datasets in online repositories, which exist in a variety of different forms at different institutional levels, ranging from university to discipline-oriented and national data infrastructures. Clearly, the number of HDCs that could be upcycled is potentially huge, but it is almost impossible to obtain an estimate of their number.

2.5. Examples

In the remainder of the paper, I refer to a selection of four HDCs from the disciplines of maritime history, an interdisciplinary field of research that combines aspects of economic, social and global history (e.g. Harlaftis, 2020), and historical metrology, an important (but underrated) auxiliary science of history and object of study from a history of science perspective (e.g. Kramper, 2019) (see table 1). They will support for my argument. Two pre-1950 HDCs are included to underline that data collection and computation are by no means recent innovations related to the advent of the computer. Two more recent HDCs allow gaining insight into the development of computational methods during the early phase of using computers for the collection, analysis and dissemination of historical data. Two HDCs are reference works; two are legacy collections.

<<insert table 1 here>>

The oldest HDC is that of Christian and Friedrich Noback's "comprehensive pocketbook" of weights and measures (Noback & Noback, 1851). On just over 1,900 pages, this book provides detailed information about everything a merchant had to know to conduct his business (Denzel, 2002). Even today, researchers consult works like the "Noback" to look up the metric equivalents of pre-modern units of measure, but they tend to ignore that the book remains vague and often silent about the sources of this information. Ronald Zupko's *Dictionary of Weights and Measures for the British Isles* (Zupko, 1985) is the more recent counterpart of the "Noback". Zupko traces mentions of weights and measures in a large number of primary sources. Although created with the assistance of a computer, the dictionary was published as a typewritten manuscript, which is now available online in various digital formats.[6] The third HDC emerged roughly in the same period as Zupko's computer-assisted

[6] https://archive.org/details/bub_gb_0l_k-XMIiQIC/mode/2up

work on historical metrology. Between 1975 and 1990, a group of British historians created a *Computerized Edition of the Liverpool Plantation Registers, 1744-1786* (Richardson, Beedham & Schofield, 1992). In 2007, the UK data service (or its predecessors) converted the data files into a platform-independent format under the name *Liverpool Shipping and Trade, 1744-1786* (further: LST).[7] The creators of LST were inspired by the Danish historian Hans Christian Johansen. In the form of an electronic database of the Danish Sound toll registers (STR) for the late eighteenth and early nineteenth centuries, Johansen (1983) aimed to produce a follow-up of the so-called Sound Toll Tables (short: STT; original Danish: *Tabeller over skibsfart og varetransport gennem Øresund*) (Bang & Korst, 1906-1953). This monumental work in seven volumes was created in the first half of the twentieth century. Alongside the LST, the STT are our second legacy collection. The STT contain about 3,000 pages of statistics about ship movements and commodity flows between the North Sea and the Baltic from 1497 to 1783. Despite substantial source criticism (e.g. Jeannin 1964; Gøbel 2010; ***), the STT have had enormous significance for the study of early modern European transport and trade (Veluwenkamp *et al.*, 2021, pp.148-152). Between 2009 and 2020, an electronic database of ships' passages through the Danish Sound, including the cargoes carried and taxes paid, was created in the Sound Toll Registers Online project (short: STRO).[8] This electronic database links entries directly to the corresponding scans of the register pages where the entry was found.

STRO readily shows how the results of digitization projects carried out in recent years may be employed to further our understanding of previous historical research methods. Alongside the electronic database, the entire archival collections that had been processed manually to create the STT has been made available as well. This gives us two reasons for including the STT as an exemplar. Firstly, there is a claim, or rather a belief, that STRO replaces the STT and overcomes its limitations (see Veluwenkamp *et al.*, 2021). While this may be true from an historian's perspective, the claim is unwarranted from an epistemological perspective. Even if the STT are 'outdated' by modern standards, they still are an exceptional computational effort. Remarkably, little is known about the creation of the STT. As an exemplar of an HDC created in the time preceding the invention of computers, the STT are a highly relevant object of historical inquiry. Thanks to STRO, a massive amount of digital materials is now available to support research on the epistemology of the STT.

2.6. A dialogue with the first retrodigitization

[7] https://doi.org/10.5255/UKDA-SN-2923-1
[8] www.soundtoll.nl

Each case in this paper should be understood as a dialogue with the accomplishments and limitations of the first mass retrodigitization of the last 20 years. All cases benefit from the availability of digital materials related to the HDC, such as scans of archival documents, historiography or OCR'ed texts. At the same time, the quality of the scans of the selected historical reference works is often insufficient, which results in 'noisy' OCR. For the available scans of the printed legacy collection STT there are some additional problems. The middle of the table was consistently lost during the scanning process. Moreover, the OCR that was automatically produced for the work cannot be used, because it does not account for the tabular structure of the source. In the case of the LST, the files could be opened and read after some trial and error, but they were almost incomprehensible due to deficiencies in their structure and the abundant use of abbreviations. In all cases, part of the work done in the first retrodigitization needs to be revised or at least requires significant post-processing when novel research questions are to be answered based on their contents.

For example, tests performed with the "Noback" have shown that the OCR quality is sufficient for answering only a few basic questions that are essentially limited to ad-hoc fact-checking (***). Although the quality of the OCR is much less of a problem for Zupko's dictionary, it suffers from the same issues. Novel insights about the composition of past metrological systems can only be obtained through comparative analysis of a large number of articles in the book, for example, all articles about places belonging to the same economic region. But scrolling through the scans or searching in the OCR'ed text are inefficient, if not impossible, methods for conducting this kind of analysis. Even though historical reference works, such as the "Noback" or Zupko's dictionary, are available in digital form, it requires upcycling rather than a bunch of scans and automatically generated OCR to preserve the historical knowledge they contain in a findable, accessible, interoperable and reusable way.

3. The issue of un-FAIR-ness

3.1. The FAIR principles and historical data collections

Across all sub-disciplines of the historical sciences and related areas of historical inquiry, many HDCs exist that could be upcycled. Yet often, it is not a sinecure, and sometimes even outright impossible to *Find* these HDCs. Focusing on research data that was produced at universities, the Swedish archivist Jakobsson (2021) laments that "[r]esearch data created within universities often ended up in a box on a bookshelf, despite university and researcher requirements to archive and share the data. (...) In addition, university researchers have typically considered such data to be their private property, to be used only for their own publications." Such HDCs, if they have survived, are found by chance, through personal contacts, or thanks to traces and hints left in historiography. Indeed, the search for HDCs often involves serendipity (Martin & Quan-Haase, 2017).

HDCs are hard, and sometimes impossible, to *Access*, most often because of the format in which the data are stored. Some HDCs are only available in print, mostly in thick volumes of tabular data, like the STT. Other HDCs are kept on index cards, punched cards, magnetic tapes, floppy disks, microfiches, or CD-ROMS. Each of these storage media is either outdated, or inefficient, or both, and hinders access to the data in its own specific ways.

HDCs are hard to make *Interoperable*, because of design choices, incompatibility of (proprietary) data formats, and lack of documentation about the collection.

These issues are widely known and discussed among librarians, archivists and historians (Rosenzweig, 2003). Initiatives promoting the production of FAIR historical data ~~in accordance with the FAIR principles~~ have become widespread since 2016, when the FAIR principles were first published (Wilkinson *et al.*, 2016). Recent work on making historical data ~~'~~FAIR~~'~~ has focused on the issue of interoperability, which is much harder to achieve and implement than the more 'technical' recommendations about making data findable, accessible and reusable with the help of persistent identifiers, metadata, data access rules and user licenses (Beretta, 2021).

Finally, HDCs were often conceived without anticipating an interest from future generations of researchers (Barats *et al.*, 2020; Kelly *et al.*, 2022), and were distributed in print or – since the 1960s – in electronic formats that more or less severely limit their *Reuse*. Often they are poorly documented and hard to understand for outsiders. As a result, there is a risk that access to the HDCs is eventually lost.

3.2. Why bother?

So why bother? Clearly, it is hard, and sometimes even impossible, to *Reuse* HDCs let alone make them *Interoperable*. They do not comply with the principles of FAIR data (cf. Wilkinson *et al.*, 2016), often suffer from methodological deficiencies and technical hindrances (cf. Huvila, 2022, p.30), and are "(...) embedded in a local context, which makes it difficult for reusers to understand and trust the data" (Faniel & Zimmerman, 2011, p.60). Reciprocally, "[t]he same factors that make it hard for scientists to reuse data collected by those from a different community also make it difficult for data producers to share data" (Faniel & Zimmerman, 2011, p.61). No wonder that "data reuse is not (…) as widespread as it might be imagined (...)" (Huvila, 2022, p.30). Yet at the same time, "[t]here is almost universal agreement that scientific data should be shared for use beyond the purposes for which they were initially collected" (Faniel & Zimmerman, 2011, p.59) and within the digital humanities there have been repeated calls to embrace the 'data deluge' (e.g. Guldi & Armitage, 2014). Moreover, in recent years, researchers have raised "(...) legitimate concerns of the present, often limited consideration of the usefulness, limits and contextuality of the currently available

and (re)used research data" (Huvila, 2022, p.30; see also: Faniel & Zimmerman, 2011; Bevan, 2015; De Freitas, 2017; Smith, 2022).

We have to bother. Deliberately throwing unFAIR HDCs out, or ignoring them, because they do not comply with current standards disregards the time and effort that had already been invested in their production. Doing the same work all over again with novel tools and methods is likely to be more expensive than reappraising what had already been achieved. More importantly, perhaps, choosing not to use existing HDCs because they are unFAIR would result in losing access to the experiences and knowledge accumulated by previous generations of scientists. This, in turn, would stand in the way of "capitalisation on existing structures and scientific legacy to avoid the duplication of means, supports and tools" (Barats *et al.*, 2020). Dismissing unfair HDCs would also make current research practices more vulnerable as it increases the risk of making the same mistakes again (Russell & Vinsel, 2018). Most importantly, however, unFAIR HDCs are essential for the systematic preservation of past methodologies for data collection and information creation.

For a very long time, it seems that the acts of selecting, collecting, and grouping data, and of making calculations, in one word: computing, were seen as something straightforward, simple, theory-less and purely pragmatic that had its place in introductions to quantitative methods for historians, or manuals for constructing historical databases (e.g. Harvey & Press, 1996; Cameron & Richardson, 2005). Fierce debates between proponents and sceptics of computational methods accompanied the early adoption of computers in historical research, but the discussion evolved around their utility *per se* rather than around the methodological implications of their use (Thomas, 2004; Greenstein, 1989). Even today, general textbooks on methods for historical research pay limited attention to computation, and if they do, they stick to the pragmatics of using computers for historical research and refrain from exploiting its methodological implications (e.g. Gunn & Faire, 2016).

4. The pursuit of scientific value gains

By targeting unFAIR HDCs, upcycling pursues two kinds of scientific value gains that directly address this lacuna in historical methodology. The first way to achieve scientific value gains is through *data rescue*, or the process of making HDCs FAIR and thus available to all. The second way is through *documentation and analysis* of the *processing steps* necessary to make HDCs FAIR. As will be discussed later, in the pursuit of these scientific value gains, the discipline of digital history unveils its paradigmatic character.

<<insert figure 1 here. Caption: Value gains from data rescue>>

Insofar as HDCs are only available in print or as data files stored on magnetic tapes, diskettes, or someone's computer, *data rescue* is primarily a technical process coordinated by memory institutions around the world. For the purpose of upcycling, however, generating or restoring access to HDCs is barely enough. Indeed, "(...) a loss of intelligibility – that is an understanding of the context in which data are generated and interpreted – is as critical as the technical or material loss of data, if not more" (Barats *et al.*, 2020). The goal of data rescue is to improve access through assessment of the design, structure and contents of the HDC and facilitate the HDC's subsequent reconfiguration in accordance with current data management standards (see figure 1). In that sense, the availability of images of books or data files in online repositories is the starting point rather than the outcome of data rescue.

From the point of view of historical methodology, *documentation and analysis* of the *processing steps* in upcycling workflows are where the most significant scientific value gains can be achieved. Meticulous documentation of upcycling workflows is deemed to reveal the technical, methodological, interpretative and creative processes, including their flaws and deficiencies, that led to the creation of an HDC. Documentation is bound to bring historical information creation processes (further: ICPs) to the surface. It puts the interactions of previous generations of researchers with primary sources in the spotlight, asking where, why and how they collected historical data; how available technical means might have affected their decisions; how their worldviews might have influenced data selection; and - last but not least - how they decided upon the format for their HDC. Upcycling thus entails a shift of focus from the results of ICPs to the actions and cognitive processes that precede and accompany them.

This shift of focus aligns well with recent developments in the study of book history. Previously, much effort has been devoted to understanding the *results* of book production and knowledge distribution in the context of grand narratives about modernization, the knowledge economy, and the rise of the West. More recently, however, researchers have started to focus on the actions and decisions that *precede, support and impact* the emergence of texts, data and information. For example, a recent discussion about the sustainability of digital data urged to "[make] the researcher's 'toolbox' readable and [explain] her/his 'ways of doing things' (…)", arguing that "(...) aspects that can be understated, obscured or considered of little value in the production of knowledge (...) could (...) play a role in reinforcing the reflective approach that is central to research" (Barats *et al.*, 2020). A comparable interest in ICPs can also be found in a variety of other disciplines, most importantly literary studies with the concept of *paratext* (Genette & MacLean, 1991; Genette, 1997; Collins & Skover, 1992, 2010; Ciotti & Lin, 2016; Skare, 2021) and information science with the concept of *paradata* (Börjesson, Skold & Huvila, 2020; Huvila, 2022). Both concepts have been adopted in the upcycling workflow that will be described ~~will be explained~~ in more detail below. Although their similarity may sound confusing

at first, both concepts play distinct roles within the upcycling workflow. They address and document different information creation processes (see section 6 for more detail). The paratext focuses on the documentation surrounding an unFAIR HDC and uses this documentation to gain insight into the 'initial' HDC. Dealing with the paratext is a preparatory step in the upcycling workflow. Contrastingly, paradata is produced during the upcycling process and documents changes to the HDC to make it FAIR. In applying these concepts, uUpcycling ~~applies these concepts to HDCs and thus~~ aims to contribute to historical information science (Boonstra *et al.*, 2006; Thaller, 2017), and in particular to the history of computational methods.

5. The upcycling workflow

5.1. Buildings blocks

In the previous sections, I have made a case for upcycling HDCs. I have argued that the HDCs themselves, the community of historians, and the scientific community in general might benefit from it. In the next sections, I introduce the building blocks of an upcycling workflow, and explain how they contribute to achieving scientific value gains. My proposed upcycling workflow consists of three parts: input, processing and documentation, and output (see figure 2).

Input consists of the unFAIR HDC and its accompanying paratext.~~,~~ The paratext comprises textual, iconic, material and factual elements that surround and give meaning to the HDC~~. and aim at making it comprehensible, like which may be understood broadly as surrounding information, like the table of contents, indices, user guides, etc.~~ (see section 5.2. for more detail) ~~that aim at making the HDC comprehensible~~. The paratext supports the main processes of the upcycling workflow: data rescue, assessment and reconfiguration. These processes go hand in hand with two tasks: (1) formulating research questions that could only be answered with better access to the data in the HDC and (2) conceptualising an 'ideal' data model for the upcycled HDC. Reconnecting the unFAIR HDC with its original 'raw' data might enhance the main upcycling processes by providing a means to verify, trace and assess the actions and decisions that were made by the creators of the HDC.

Data rescue might have to start with technical procedures to generate or restore access to HDCs, but they are not our main concern. Data rescue is understood here as part of an iterative process to improve access to an unFAIR HDC and make it interoperable and reusable. Starting point of data rescue are the scans or data files of the unFAIR HDCs as well as the paratext that documents their structure and contents. Insofar as the paratext is concerned, data rescue secures the availability in digital format of parts of the paratext that can significantly improve access to and use of the HDC. Assessment examines what needs to be done to reconfigure the unFAIR HDC and make it FAIR. Answers to this question are sought through manual and digital exploration of the HDC, which is supported by comparison

of the HDC with the paratext and reconnected primary source documents. The goal of assessment is to identify and remove errors and omissions in the data model of the HDC. In doing so, it explicates how the data model can be reconfigured and why this should be done. Removing data modelling errors effectively means changing the data model and thus reconfiguring the HDC.

Paradata documents the implementation of the iterative processes of data rescue, assessment and reconfiguration, and captures the essential steps of data rescue, assessment and reconfiguration in a systematic way. The goal of paradata capturing is to facilitate the subsequent analysis of the ICPs related to the HDC (see section 6 for more detail). The output of upcycling is not a single product, like a book, or an article, or an electronic dataset, but an upcycled HDC, which comprises the 'original' HDC, the paradata produced during upcycling, the 'new' HDC and its reconnection to the original 'raw' data. Reconfigured in that way, the upcycled HDC serves as a starting point for the study of historical ICPs using paradata analysis.

<<insert figure 2 here. Caption: The upcycling workflow>>

5.2. Using the paratext

Upcycling uses the paratext found in documentation, guides, descriptions, prefaces, indexes and the HDC itself. The concept of paratext as introduced by the French literary scholar Gérard Genette (1930-2018) targets the information that surrounds the actual text of a book, starting with advertisements about forthcoming publication, title page, front and back cover, and ending with prefaces, indexes and reviews (Genette, 1997). According to Genette, each element of the paratext of a book provides insight into the creation of the book and contributes to understanding its contents. Thus, the paratext is "(...) the means by which a text makes a book of itself and proposes itself as such to its readers, and more generally to the public" (Genette & MacLean, 1991, p.261). Developed in the realm of literary studies, the concept of paratext focuses heavily on books (Ciotti & Lin, 2016, p.VI-VII), although its applicability is not limited to literary works (e.g. Birke & Christ, 2013; Skare, 2020, 2021). Therefore, Ciotti and Lin suggested defining paratext more broadly as a reflection of "the activities of everyone involved in the production, transmission, dissemination and reception of the manuscript and its content: authors, editors, scribes, artisans, commentators, readers, sellers, owners and so on…" (Ciotti & Lin, 2016, p.VIII). Significantly, the paratext also comprises iconic, material and factual elements, such as illustrations, the typographical features of a work, or the author's gender, age and reputation (Skare 2020, p.512), which target, among other things, the HDC's physical form. These elements contribute to overcoming the limitations of a one-sided hermeneutic approach that focuses on the immaterial and contribute to the renewed interest

in materiality in the digital age (for more information, see Skare, 2021). A comprehensive analysis of the paratext of an HDC sheds light on the intentions, methods, creative decisions and implementation strategies of the creators of the collection and is therefore a prerequisite for upcycling. Analysis of the paratext is essentially an inductive ~~hermeneutic~~ process (Genette & MacLean, 1991, pp.270-271) that is similar to the methods of historical praxeology, the study of practices in the past (Haasis & Rieske, 2015; Füssel, 2022).

In the "Noback", Zupko's dictionary, and the STT, elements of the paratext are found in the works themselves, before and after their 'core'. In the 'Vorwort' (*Eng.* preface) and 'Einleitung' (*Eng.* introduction) (Noback & Noback, 1851, pp.V-LII), the 'Introduction' (Zupko, 1985, pp.x-xiii) or the 'Vorord / Avant-Propos' (Bang & Korst 1906, pp.III-X; 1922, pp.III-X; 1930, pp.VII-XX), the authors explain their motives and describe, in more or less detail, how they collected and systematized their data. Additionally, depending on the example, one or more indexes are included for geographical locations, commodities and abbreviations. Alongside the introduction and acknowledgements, the paratext of Zupko's dictionary also consists of several auxiliary tables, e.g. for abbreviations (Zupko, 1985, pp.xxii-xxiv), which precede the dictionary, and an extensive bibliography, which is added at the end (Zupko, 1985, pp.447-520). In the LST, a separate 'guide' accompanies the data files uploaded to the UK data archive. This 'guide' describes a relational database constructed in *dBase IV*, explains the elements of each table, and provides 'codenames for occupations of owners of vessels', 'codenames for places included in database' and abbreviations of the 'sources used in the database' as appendices to the guide (Richardson *et al.*, 1992). By processing parts of the paratext, access to the HDC is easily improved. For example, the appendices in the 'guide' to the database of the Liverpool Plantation Registers can be processed with an OCR tool to make them machine-readable and linkable to the data files in the collection.

5.3. Adopting an emic perspective

The elements of the paratext that can be found in the immediate neighbourhood of the 'core' of the HDC provide valuable support for understanding the intricacies of their composition, but based on these elements alone, a comprehensive understanding of the HDC is hard to attain. Where the documentation is insufficient, exploratory analysis of the HDC itself should be conducted to fill remaining gaps in the documentation (cf. Ciotti & Lin, 2016, p.VIII). The extent of such posthumous paratext creation varies. In some cases, very little documentation is available; in other cases, the documentation is abundant and raises only minor questions.

Posthumous paratext creation, or the attempt to expand and improve the available documentation by means of exploratory analysis of the unFAIR HDC, is challenging. It requires 'diving into' the source and thus supports the processes of data rescue, assessment and reconfiguration. Posthumous paratext creation as a means to fill in essential gaps in the

documentation requires looking at unFAIR HDCs from the perspective of the human agents that were involved in its creation. In other words, it requires an emic perspective towards historical information creation processes.

The mass digitization pursued at memory institutions around the world offers profoundly new possibilities for adopting an emic perspective. In many cases, the HDC can be reconnected to scans of archival sources or library collections. For example, the extensive bibliography at the end of Zupko's dictionary could be used as a starting point for tracing online versions of these works, which could then be connected to the dictionary itself. In the case of the STT and LST, scans of the original archival sources are readily available on dedicated websites.[9] These scans allow reiterating the data creation processes of previous generations of researchers, tracing the missing elements in the paratext, and improving our understanding of the processes of seeking, collecting, selecting and systematising historical data. The example of the "Noback", however, shows that reconnection cannot be guaranteed.

The potential benefits of reconnecting the unFAIR HDC to scans of archival documents and books for the purpose of an emic perspective towards historical information creation processes are high, but the complexity of the process should not be underestimated. Reconnection requires careful preparation and should only be carried out if significant value gains can be expected. It adds an extra dimension to the upcycling workflow, but it also comes with additional technical requirements.

5.4. The scenario-based method

The processes of assessment and reconfiguration employ techniques from the realms of automated text recognition, natural language processing and data modelling (see figure 1). Before embarking upon these processes, it is necessary to formulate possible research questions and conceptualise an 'ideal' data model for the upcycled HDCs (see figure 3). The process of anticipating research questions for the benefit of streamlining data modelling efforts is known in knowledge engineering as the scenario-based method. It relates the planning process to so-called competency questions, which are questions that specialist users of the source would like to examine and answer with the help of a data model (Kendall & McGuinness, 2019). Lodi *et al.* (2017) and Carriero *et al.* (2021) have successfully adapted this renowned scenario-based method to questions of digital cultural heritage. However, addressing the processed data in HDCs requires modelling approaches that are distinct from those that target the 'data about data', or metadata, in digital heritage collections (***).

[9] www.soundtoll.nl; https://microform.digital/boa/series/20/the-transatlantic-slave-trade-1675-1907; www.slavevoyages.org

<<insert figure 3 here. Caption: Applications of the scenario-based method

For the “Noback”, we assume an upcycled HDC that could be described as a ‘Web of Science’, in which the contents of the “Noback” are made fit for text reuse detection in earlier (pre-1850) metrological reference works. This will make it possible to tackle research questions about the origins of the data compiled by Christian and Friedrich Noback. For Zupko’s dictionary, we assume an upcycled HDC that could be described as hierarchical ‘Tree of Knowledge’. This will allow answering questions about long-term processes of accumulation and dissemination of metrological knowledge as well as about the long-term development of metrological systems on the British Isles. For the LST, we anticipate a normalized relational data model with links to original archival documents (where available) as the result of upcycling. This would provide a novel basis for answering enduring questions about ownership relations in the shipping industry and the role of Liverpool in the trans-Atlantic slave trade in the eighteenth century. Finally, for the STT, we assume an upcycled HDC that links the semantic data in the STT to the electronic database of the Sound toll registers online (STRO), and thus also to the scans of the registers. This combination of sources makes it possible to develop a kind of ‘genetic criticism’ for HDCs (cf. Van Hulle, 2022). Rather than focusing on the published statistics *per se* to reach a verdict about their reliability, the upcycling workflow allows tracing the cognitive processes and pragmatic decisions made by the compilers of the STT in an unprecedented way.

6. Paradata production and analysis

The previous section did not provide lengthy descriptions of all the different steps that were (or could be) taken to improve access to and use of unFAIR HDCs, but instead limited itself to a few examples showing what data rescue, assessment and reconfiguration entail in practice. These few examples are deemed sufficient for a discussion of the production of paradata as a basic requirement for gaining insight into the historical ICPs that are ‘hidden’ in each HDC.

6.1. The concept of paradata

Paradata is a well-known concept in survey research and archaeology (e.g. Bentkowska-Kafel & Denard, 2012; Huvila, 2022). It recently started to attract the attention of digital humanists as well, in particular in relation to the issue of reproducibility of research results (Schubert, 2021; Schöch *et al.*, 2020). Isto Huvila summarises the importance of paradata (and more generally of the study of information creation processes) as follows:

> “Without documentation of how data was captured or created, in what types of conditions, what tools were used, who did what and when, what qualifications and

> experience the individuals had, and for instance, who made decisions and on what premises it is impossible to know the extents of the data, what was left out and why, and in general, how well the data suits for addressing other types of research questions than the original ones, and what needs to be taken into account when using it." (Huvila, 2022, p.31).

Whereas the need for and usefulness of paradata is undisputed and various technical solutions for capturing paradata are available (for an overview, see Huvila *et al.*, 2021; Börjesson *et al.*, 2020), a methodology for producing paradata is still in its infancy (Huvila 2022). Recurring questions related to the use of paradata include: What information can function as paradata? Where can it be found? And how much capturing is enough? (Huvila, 2022).

6.2. Capturing paradata

Paradata is captured in the form of narrative statements describing each action in a machine-readable and exchangeable format. However diverse and individual these narrative statements may seem, they fall within a certain range of viable options and share certain characteristics. In many cases, capturing is informed by the paratext as 'contextual information provider'. Capturing paradata relies on a vocabulary and a set of rules that need to be elaborated (Huvila, 2022). In the upcycling workflow, paradata has a double scope.

On the one hand, paradata captures information about the methods employed by researchers in the past, i.e. information about the behaviour of previous generations of historians towards the raw data that they worked with. The paratext delivered with, or produced posthumously about the unFAIR HDC, informs the rescue, assessment and reconfiguration stages of upcycling and thus becomes part of the paradata.

On the other hand, paradata also captures information about the methods employed in the present, i.e. information about the steps taken towards data rescue, assessment and reconfiguration. The latter kind of paradata is of essential importance for the reproducibility of the upcycling workflow. At least in part, it can be captured automatically, whereas the former kind requires a vocabulary, or a description language, and description rules to facilitate paradata analysis.

6.3. Analysing paradata

Upcycled HDCs comprise the 'original' HDC, the paradata produced during the iterative process of upcycling, the 'new' HDC and - if available - links to the original 'raw' data (see figure 2). Having been reconfigured in that way, upcycled HDCs may now serve as starting points for the study of historical ICPs using paradata analysis. Inspired by Gardin's logicist programme, a long-term project developed in archaeology since the 1970s (Gardin, 1999;

Marlet *et al.*, 2019; Buard *et.al.*, 2020), paradata analysis aims at identifying the cognitive structure of HDCs.

Paradata analysis studies the actions undertaken to rescue, assess and reconfigure unFAIR HDCs insofar as they had been captured during upcycling. The goal of paradata analysis is to further our understanding of historical information creation processes. To achieve this goal, paradata analysis relies on upcycled HDCs, which comprise both the 'original' unFAIR HDC, the documented upcycling process and the 'new' HDC. Therefore, upcycled HDCs can provide novel insights into two different ICPs (see figure 4).

<<insert figure 4 here. Caption: The study of information creation processes using paradata analysis>>

Starting from the unFAIR HDC, the first ICP documents how data were rescued and made fit for reuse. The first ICP comprises the workflows that digital historians and data scientists develop and implement to achieve an upcycled HDC. The second ICP documents how 'raw data' from primary sources was put into an HDC by previous generations of scientists. The second ICP targets their workflows. In principle, at least part of the information for the second ICP is found in the paratext of HDCs. However, even today, the practice of documenting decisions made during the process of collecting data is not a priority in the historical sciences, and the further we go back, the poorer the available documentation becomes. Therefore, the second ICP also entails producing a posthumous paratext during upcycling (see above). Thus, while rescuing an HDC and preparing it for reuse, valuable insights can be obtained about the methods applied by previous generations of scientists.

Paradata analysis can be understood as a condensation process that aims at unveiling the logico-semantic structure of data rescue, assessment and reconfiguration procedures implemented by 'upcyclers', on the one hand, and of data collection and systematisation procedures of previous generations, on the other hand. The result is a tree structure in which propositions (narrative statements) about the actions undertaken with the data are condensed into more general procedures, practices, methods and – eventually – cognitive processes. In doing so, paradata analysis sheds light on reasoning processes in the field of historical information science and contributes to the study of the historical methods that were - and still are – applied to 'raw' historical data sources and 'processed' historical data collections. Ultimately, the interpretation of these reasoning processes is the task of the digital historian, who - through his analysis of ICPs – contributes to the history of computational methods.

7. Discussion

My proposed upcycling workflow consists of three parts: input, processing and documentation,

and output. Input consists of the unFAIR historical data collection and its accompanying paratext, which informs and supports the main data rescue processes: exploration, assessment and reconfiguration. Reconnection of the unFAIR historical data collection with its original 'raw' data supports these iterative processes. The implementation of these processes is captured in paradata. The output of upcycling is not a single product, like a book, or an article, or an electronic dataset, but an upcycled historical data collection, which comprises the 'original' collection, the paradata produced during upcycling, the 'new' collection and its reconnection to the original 'raw' data. Reconfigured in that way, the upcycled historical data collection serves as the starting point for the study of historical information creation processes based on paradata analysis. The latter entails a shift of focus from the results of ICPs, which we can find in historical data collections, to the cognitive processes that precede and accompany their creation. In the pursuit of this shift of focus, upcycling unveils its paradigmatic character.

Upcycling's first contribution is to digital history, in general, and to the history of computational methods, in particular. Adopting a long-term view that also includes HDCs that precede the advent of computers, the study of historical information creation processes furthers our understanding of computational methods and their role in historical research. Upcycling in general, and paradata analysis in particular, address the intimate relationship between the history of computing and developments in historical methods, and thus contribute to a better understanding of the 'early history', or should we say the 'pre-history', of digital history. Indeed, through analysis of ICPs in HDCs, I hope to draw attention to two elements in the relationship between historians and computers that seem to have fallen off the radar in recent years. The first element is that of the continuities between 'analog' and 'digital' methods of historical research. The second element is that of the parallel development and application of computational methods in the humanities, on one hand, and the historical social sciences, on the other. Indeed, as early as 2011, the late Peter Haber already pointed out that the parallel development of computer-assisted research using historical data collections seems to have been lost in recent accounts of the history of 'digital humanities' and its predecessors (Haber, 2011).

Upcycling's second contribution is to the emerging field of digital hermeneutics. Upcycling responds to Andreas Fickers' call for an 'Update for Hermeneutics' in the form of a praxeological reflection on the impact of digital Infrastructure, tools and sources on the practice of historical research (Fickers, 2020, p.2; Fickers *et al.*, 2022). Through its focus on the ICPs that precede and accompany historical research, upcycling contributes in a specific and hitherto under-exploited way to historical data criticism as an essential part of digital hermeneutics (Fickers, 2020, pp.6-7). In particular, upcycling suggests that ICPs are an important category in the practice of digital historians. Often, however, ICPs remain implicit or

are subsumed under the more general heading of 'data collection' in current data management plans (see e.g. Barats *et al.*, 2020). This, in turn, has a negative impact on scientific transparency, interoperability and, ultimately, sustainability of HDCs (Börjesson *et al.*, 2020).

8. Conclusions

Upcycling is an agenda for research in the field of digital history. In the preceding sections, the constituent parts of this agenda have been outlined. The next step should be to start implementing upcycling workflows in order to test and improve its conceptualization. To do so, it is of vital importance that the selected use cases are elaborated further. Moreover, it is essential to start working on a vocabulary and description rules for paradata capturing and analysis. And last but not least, following the example set by the National Agricultural Library in the US (Clarke & Shiue, 2020; Shiue *et al.*, 2021; see also Faniel & Zimmerman, 2011, pp.66-67), it is essential to start working on an assessment framework for selecting or not-selecting historical data collections for upcycling based on clearly defined feasibility requirements and expected value gains.

References

Avesani, P., *et al.* (2019), "The open diffusion data derivatives, brain data upcycling via integrated publishing of derivatives and reproducible open cloud services", *Scientific Data*, Vol. 69 No. 6, pp.1-13. https://doi.org/10.1038/s41597-019-0073-y

Bang, N.E. & Korst, K. (Ed.s), *Tabeller over skibsfart og Varetransport gennem Øresund / Tables de la navigation et du transport des marchandises passant par le Sund 1497–1783.* Gyldendalske Boghandel – Nordisk Verlag, København / Otto Harrassowitz, Leipzig, 1906-1953.

Barats, C., Schafer, V., & Fickers, A. (2020), "Fading Away... The challenge of sustainability in digital studies", *Digital Humanities Quarterly*, Vol. 14 No. 3. https://www.proquest.com/scholarly-journals/fading-away-challenge-sustainability-digital/docview/2553557106/se-2

Baten, J., & Van Zanden, J.L. (2008), "Book production and the onset of modern economic growth", *Journal of Economic Growth*, Vol. 13 No. 3, pp.217-235.

Bellingradt, D., & Salman, J. (2017), "Books and book history in motion: Materiality, sociality and spatiality", Bellingradt, D., Nelles, P., & Salman, J. (Ed.s), *Books in Motion in Early Modern Europe: Beyond production, circulation and consumption*. Palgrave Macmillan, Cham, pp.1-11.

Bentkowska-Kafel, A., Denard, H., & Baker, D. (Ed.s) (2012), *Paradata and transparency in virtual heritage*, Ashgate, Farnham.

Beretta, F. (2021), "A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME)", *Semantic Web*, Vol. 12 No. 2, pp.279-294. DOI: 10.3233/SW-200416

Berg, M. (2007), "The Genesis of 'Useful Knowledge'", *History of Science*, Vol. 45 No. 2, pp. 123–133. https://doi.org/10.1177/007327530704500201

Besomi, D. (Ed.) (2012), *Crises and Cycles in Economics Dictionaries and Encyclopedias*, Routledge, Abingdon / New York.

Bevan, A. (2015), "The data deluge", *Antiquity,* Vol. 89 No. 348, pp.1473-1484. doi:10.15184/aqy.2015.102

Birke, D., & Christ, B. (2013), "Paratext and Digitized Narrative: Mapping the Field", *Narrative*, Vol. 21 No. 1, pp.65-87. doi:10.1353/nar.2013.0003

Boonstra, O., Breure, L., & Doorn, P. (Ed.s) (2006), *Past, present and future of historical information science*, DANS. 10.26530/OAPEN_353255

Börjesson, L., Sköld, O., & Huvila, I. (2020), "Paradata in Documentation Standards and Recommendations for Digital Archaeological Visualisations", *Digital Culture & Society*, Vol. 6 No. 2, pp.191-220. https://doi.org/10.14361/dcs-2020-0210

Buard, P.Y., Zadora-Rio, E., Chauveau, J., Roger, J., & Marlet, O. (2020), *Publishing an Archeological Excavation Report in a Logicist Workflow*. http://dx.doi.org/10.15496/publikation-43222

Buringh, E., & Van Zanden, J.L. (2009), "Charting the "Rise of the West": Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries", *The Journal of Economic History,* Vol. 69 No. 2, pp.409-445. doi:10.1017/S0022050709000837

Cameron, S. & Richardson, S. (2005), *Using computers in history*, Palgrave Macmillan, Basingstoke / New York, 2005.

Carriero, V.A., Gangemi, A., Mancinelli, M.L., Nuzzolese, A.G., Presutti, V., & Veninata, C. (2021), "Pattern-based Design Applied to Cultural Heritage Knowledge Graphs", *Semantic Web*, Vol. 12, pp.313 – 357. 10.3233/SW-200422

Cherchi Usai, P. (2010), "The Conservation of Moving Images", *Studies in Conservation*, Vol. 55, pp. 250-257, DOI: 10.1179/sic.2010.55.4.250

Chiron, G., Doucet, A., Coustaty, M., Visani, M. & Moreux, J.-P. (2017), "Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information", *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp.1-4, doi: 10.1109/JCDL.2017.7991582

Christy, M., Gupta, A., Grumbach, E., Mandell, L., Furuta, R., & Gutierrez-Osuna, R. (2017), "Mass Digitization of Early Modern Texts With Optical Character Recognition", *Journal on Computing and Cultural Heritage*, Vol. 11 No. 1, pp.1-25. https://doi.org/10.1145/3075645

Ciotti, G. & Lin, H. (Ed.s) (2016), *Tracing Manuscripts in Time and Space through Paratexts*, De Gruyter, Berlin / Boston. https://doi.org/10.1515/9783110479010

Clarke, C.T. & Shiue, H.S.Y. (2020), *Final Report and Recommendations of the Data Rescue Project at the National Agricultural Library*. https://doi.org/10.13016/kpt7-cqgr

Collins, R., & Skover, D. (2010), "Paratexts as Praxis", *Neohelicon*, Vol. 37 No. 1, pp.33-51. https://doi.org/10.1007/s11059-010-0050-0

Cordell, R. (2017), ""Q i-jtb the Raven": Taking Dirty OCR Seriously", *Book History*, Vol. 20, pp.188-225. 10.1353/bh.2017.0006

de Freitas, E. (2017), "The temporal fabric of research methods: Posthuman social science and the digital data deluge", *Research in Education*, Vol. 98 No. 1, pp.27–43. https://doi.org/10.1177/0034523717723386

De Kramer, M. (2022), "3D models are easy. Good 3D models are not", Fickers, A., Tatarinov, J. (Ed.s), *Digital History and Hermeneutics: Between Theory and Practice*, De Gruyter Oldenbourg, Berlin / Boston, pp. 239-258. https://doi.org/10.1515/9783110723991-012

Denzel, M.A. (2002), "Handelspraktiken als wirtschaftshistorische Quellengattung vom Mittelalter bis in das frühe 20. Jahrhundert. Eine Einführung", Denzel, M.A., Hocquet, J.-C., & Witthöft, H. (Ed.s), *Kaufmannsbücher und Handelspraktiken vom Spätmittelalter bis zum beginnenden 20. Jahrhundert — Merchant's Books and Mercantile Pratiche from the Late Middle Ages to the Beginning of the 20th Century*, Steiner Verlag, Stuttgart, pp.11-45.

Dobrin, L.M. & Schwartz, S. (2021), "The social lives of linguistic legacy materials", *Language Documentation and Description*, Vol. 21, pp.1-36. http://www.elpublishing.org/PID/244

Faniel, I.M., & Zimmerman, A. (2011), "Beyond the Data Deluge. A research agenda for large-scale data sharing and reuse", *International Journal of Digital Curation*, Vol. 6 No. 1, pp.58-69. https://doi.org/10.2218/ijdc.v6i1.172

Fickers, A. (2020), "Update für die Hermeneutik. Geschichtswissenschaft auf dem Weg zur digitalen Forensik?", *Zeithistorische Forschungen–Studies in Contemporary History*, Vol. 17 No. 1, pp.157-168. https://doi.org/10.14765/zzf.dok-1765

Fickers, A., Tatarinov, J., & Heijden, T. van der (2022), "Digital history and hermeneutics – between theory and practice: An introduction", in: Fickers, A. & Tatarinov, J. (Ed.s), *Digital History and Hermeneutics. Between Theory and Practice*, De Gruyter Oldenbourg, Berlin / Boston, pp.1-19. https://doi.org/10.1515/9783110723991

Füssel, M. (2022), "Praxeologie als Methode", Haas, S. (Ed), *Handbuch Methoden der Geschichtswissenschaft*, Springer, Wiesbaden, pp.1-19. https://doi.org/10.1007/978-3-658-27798-7_18-1

Gardin, J.-C. (1999), “Archéologie, formalisation et sciences sociales”, *Sociologie et sociétés*, Vol. 31 No. 1, pp.119–127. https://doi.org/10.7202/001282ar

Genette, G. (1997). *Paratexts. Tresholds of Interpretation*. Cambridge University Press, Cambridge.

Genette, G., & Maclean, M. (1991), “Introduction to the Paratext”, *New Literary History*, Vol. 22 No. 2, pp.261–72. https://doi.org/10.2307/469037

Gøbel, E. (2010), “The Sound Toll Registers Online project, 1497–1857”, *International Journal of Maritime History*, Vol. 22 No. 2, pp.305–324.

Greenstein, D.I. (1989), “A Source-Oriented Approach to History and Computing: The Relational Database” *Historical Social Research / Historische Sozialforschung*, Vol. 14 No. 3, pp. 9–16. http://www.jstor.org/stable/20754411

Guldi, J. & Armitage, D. (2014), *The History Manifesto*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781139923880

Gunn, S. & Faire, L. (Ed.s) (2016), *Research methods for history*. Edinburgh University Press, Edinburgh.

Haasis, L. & Rieske, C. (2015), *Historische Praxeologie : Dimensionen vergangenen* Handelns, *Verlag* Ferdinand Schöningh, Paderborn.

Haber, P. (2011), *Digital Past: Geschichtswissenschaft im digitalen Zeitalter*, Oldenbourg Wissenschaftsverlag, Berlin / Boston. https://doi.org/10.1515/9783486712339

Harlaftis, G. (2020), “Maritime history: A new version of the old version and the true history of the sea”, *International Journal of Maritime History*, Vol. 32 No. 2, pp.383–402. https://doi.org/10.1177/0843871420924243

Harvey, C. & Press, J. (Ed.s) (1996), Databases in Historical Research. Palgrave, London. https://doi.org/10.1007/978-1-349-24392-1_1

Heitman, C., Worthy, M., & Plog, S. (2017), “Innovation through Large-Scale Integration of Legacy Records: Assessing the “Value Added” in Cultural Heritage Resources”, *Journal on Computing and Cultural Heritage* Vol. 10 No. 3, Article 17, 10 pages. https://doi.org/10.1145/3012288

Herr, W. (2012), “VD 16,17 und 18: Retrospektive nationalbibliographische Verzeichnung vor dem Hintergrund sich verändernder technischer Rahmenbedingungen”, *Perspektive Bibliothek* Vol. 1 No. 2, pp.106-136.

Huvila, I. (2022), "Improving the usefulness of research data with better paradata", *Open Information Science*, Vol. 6 No. 1, pp.28-48. https://doi.org/10.1515/opis-2022-0129

Huvila, I., Sköld, O. and Börjesson, L. (2021), "Documenting information making in archaeological field reports", *Journal of Documentation*, Vol. 77 No. 5, pp. 1107-1127. https://doi.org/10.1108/JD-11-2020-0188

Jacob, M.C. (2014), *The first knowledge economy: Human capital and the European economy, 1750–1850*, Cambridge University Press, Cambridge.

Jaillant, L. (2022), "Design Thinking, UX and Born-digital Archives: Solving the Problem of Dark Archives Closed to Users", Jaillant, L. (Ed.), *Archives, Access and Artificial Intelligence. Working with Born-Digital and Digitized Archival Collections*, Transcript, Bielefeld, pp. 83-107.

Jakobsson, U. (2021), "Digital Archaeological Archiving in Sweden: the Swedish National Data Service perspective", *Internet Archaeology,* Vol. 58. https://doi.org/10.11141/ia.58.18

Jeannin, P. (1964), "Les comptes du Sund comme source pour la construction de l'activité économique en Europe (XVIe–XVIIIe siècle)", *Revue historique,* Vol. 121, pp.55–102 and 307–340.

Jeannin, P., Kaiser, W. & Hoock, J. (Ed.s) (1991-2001), *Ars mercatoria. Handbücher und Traktate für den Gebrauch des Kaufmanns / Manuels et traités à l'usage des marchands, 1470-1820. Eine analytische Bibliographie*, Ferdinand Schöningh, Paderborn.

Johansen, H.C. (1983), *Shipping and Trade between the Baltic and Western Europe 1784–95*. Odense: Odense University Press.

Kelly, J.A., Farrell, S.L., Hendrickson, L.G., Luby, J. & Mastel, K.L. (2022), "A Critical Literature Review of Historic Scientific Analog Data: Uses, Successes, and Challenges", *Data Science Journal*, Vol. 21 No. 14, pp.1–11. https://doi. org/10.5334/dsj-2022-014

Kendall, E.F., & McGuinness, D.L. (2019), *Ontology engineering,* Morgan and Claypool. 10.2200/S00834ED1V01Y201802WBE018

Kramper, P. (2019), *The Battle of the Standards. Messen, Zählen und Wiegen in Westeuropa, 1660-1914*, De Gruyter, Berlin / Boston.

Krünitz, J.G. (1773-1858), *Ökonomische Enzyklopädie oder allgemeines System der Staats-, Stadt-, Haus- und Landwirthschaft, in alphabetischer Ordnung*, http://www.kruenitz1.uni-trier.de/

Lodi, G., Asprino, L., Nuzzolese, A.G., Presutti, V., Gangemi, A., Recupero, D.R., Veninata, C., & Orsini, A. (2017), "Semantic Web for Cultural Heritage Valorisation", in: Hai-Jew, S. (Ed.), *Data Analytics in Digital Humanities*. *Multimedia Systems and Applications*, Springer, Cham, pp.3-37. https://doi.org/10.1007/978-3-319-54499-1_1

MacFarland, K., & Vokes, A.W. (2016), "Dusting off the data: curating and rehabilitating archaeological legacy and orphaned collections", *Advances in Archaeological Practice*, Vol. 4 No. 2, pp. 161-175. 10.7183/2326-3768.4.2.161

Marlet, O., Zadora-Rio, E., Buard, P.-Y., Markhoff, B., & Rodier, X. (2019), "The Archaeological Excavation Report of Rigny: An Example of an Interoperable Logicist Publication" *Heritage*, Vol. 2 No. 1, pp.761-773. https://doi.org/10.3390/heritage2010049

Martin, K., & Quan-Haase, A. (2017), ""A process of controlled serendipity": An exploratory study of historians' and digital historians' experiences of serendipity in digital environments", *Proceedings of the Association for Information Science and Technology*, Vol. 54 No. 1, pp.289-297.

McCartney, L. (2018), "The Evolution and Issues Encountered in Providing Meaningful and Useful Online Access to Oral History Recordings", *International Association of Sound and Audiovisual Archives (IASA) Journal*, Vol. 47, pp. 30–40. https://doi.org/10.35320/ij.v0i47.53

McDonough, W., & Braungart, M. (2013), *The upcycle: Beyond sustainability--designing for abundance*. Macmillan.

McKenzie, D.F. (1999), *Bibliography and the Sociology of Texts*. Cambridge University Press, Cambridge.

Meroño-Peñuela, A. and Hoekstra, R. (2014), "What is Linked Historical Data?", Janowicz, K., Schlobach, S., Lambrix, P., & Hyvönen, E. (Eds.), *Knowledge engineering and knowledge management. 19th international conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, Springer, Cham, pp.282–287.

Mokyr, J. (2002), *The Gifts of Athena. Historical Origins of the Knowledge Economy*, Princeton University Press, Princeton. https://doi.org/10.1515/9781400829439

Noback, C. & Noback, F. (1851), *Vollständiges Taschenbuch der Münz-, Maass-, und Gewichtsverhältnisse, der Staatspapiere, des Wechsels- und Bankwesens, und der Usanzen aller Länder und Handelsplätze*, Brockhaus, Leipzig.

Ricci, S. (2008), "Saving, Rebuilding, or Making: Archival (Re) Constructions in Moving Image Archives", *The American Archivist*, Vol. 71, pp. 433–55. http://www.jstor.org/stable/40294525.

Richardson, D., Beedham, K., & Schofield, M.M. (1992), *Liverpool Shipping and Trade, 1744-1786. A computerised edition of the Liverpool plantation registers* (Unpublished manuscript). https://sp.ukdataservice.ac.uk/doc/2923/mrdoc/pdf/guide.pdf

Ronald K., Collins, L., & Skover, D.M. (1992), "Paratexts", *Stanford Law Review*, Vol. 44 No. 3, pp.509–52. https://doi.org/10.2307/1228974

Rosenzweig, R. (2003), "Scarcity or abundance? Preserving the past in a digital era", *The American historical review*, Vol. 108 No. 3, pp.735-762. https://doi.org/10.1086/ahr/108.3.735

Russell, Andrew L., and Lee Vinsel. "After Innovation, Turn to Maintenance." *Technology and Culture* 59, no. 1 (2018): 1-25. doi:10.1353/tech.2018.0004

***

***

Schöch, C., Van Dalen-Oskam, K., Antoniak, M., Jannidis, F., & Mimno, D. (2020), "Replication and Computational Literary Studies", *Digital Humanities Conference 2020 (DH2020), Ottawa, Canada*. https://doi.org/10.5281/zenodo.3893428

Schubert, C. (2021), "Digital Humanities auf dem Weg zu einer Wissenschaftsmethodik: Transparenz und Fehlerkultur", *Digital Classics Online* Vol. 7, pp.39-53. https://doi.org/10.11588/dco.2021.7.82371

Shiue, H.S.Y., Clarke, C.T., Shaw, M., Hoffman, K.M., & Fenlon, K. (2021), "Assessing Legacy Collections for Scientific Data Rescue", Toeppe, K., Yan, H., & Chu, S.K.W. (Ed.s), *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17-31 2011, Proceedings, Part II*, Springer, Cham, pp. 308–318. https://link.springer.com/chapter/10.1007/978-3-030-71305-8_25.

Skare, R. (2020), "Paratext", *Knowledge Organization,* Vol. 47 No. 6, pp.511-519. https://doi.org/10.5771/0943-7444-2020-6-511

Skare, R. (2021), "The paratext of digital documents", *Journal of* Documentation, Vol. 77 No. 2, pp.449–460. 10.1108/JD-06-2020-0106.

Smith, G. N. (2022), "The Promise and Peril of the Data Deluge for Historians", *Journal of Cognitive Historiography*, Vol. 6, No. 1-2, pp.277–287. https://doi.org/10.1558/jch.21156

Sommer, D. (2010), "VD16, VD17, VD18: Diversität und Integration", *ABI-Technik*, Vol. 30 No. 2, pp.120-128.

Thaller, M. (2017), "Historical Information Science: Is There Such a Thing? New Comments on an Old Idea [1993]", *Historical Social Research, Supplement*, Vol. 29, pp.260-286. https://doi.org/10.12759/hsr.suppl.29.2017.260-286

Thieberger, N. (2018), "Research methods in recording oral tradition: Choosing between evanescence of the digital and senescence of the analog", levenberg, I., Neilson, T., Rheams, D. (Ed.s), *Research Methods for the Digital Humanities*. Palgrave Macmillan, Cham, pp. 233-241. https://doi.org/10.1007/978-3-319-96713-4_13

Thomas III, William G. (2004), "Computing and the historical imagination", Schreibman, S., Siemens, R., & Unsworth, J. (Ed.s), *A companion to digital humanities*, Malden / Oxford / Carlton, Blackwell Publishing, pp.56-68. https://doi.org/10.1002/9780470999875.ch5

Van Hulle, Dirk (2022), *Genetic Criticism: Tracing Creativity in Literature*, Oxford University Press, Oxford.

Vearncombe, J., Riganti, A., Isles, D., & Bright, S. (2017), "Data upcycling", *Ore Geology Reviews*, Vol. 89, pp.887-893. https://doi.org/10.1016/j.oregeorev.2017.07.009

***

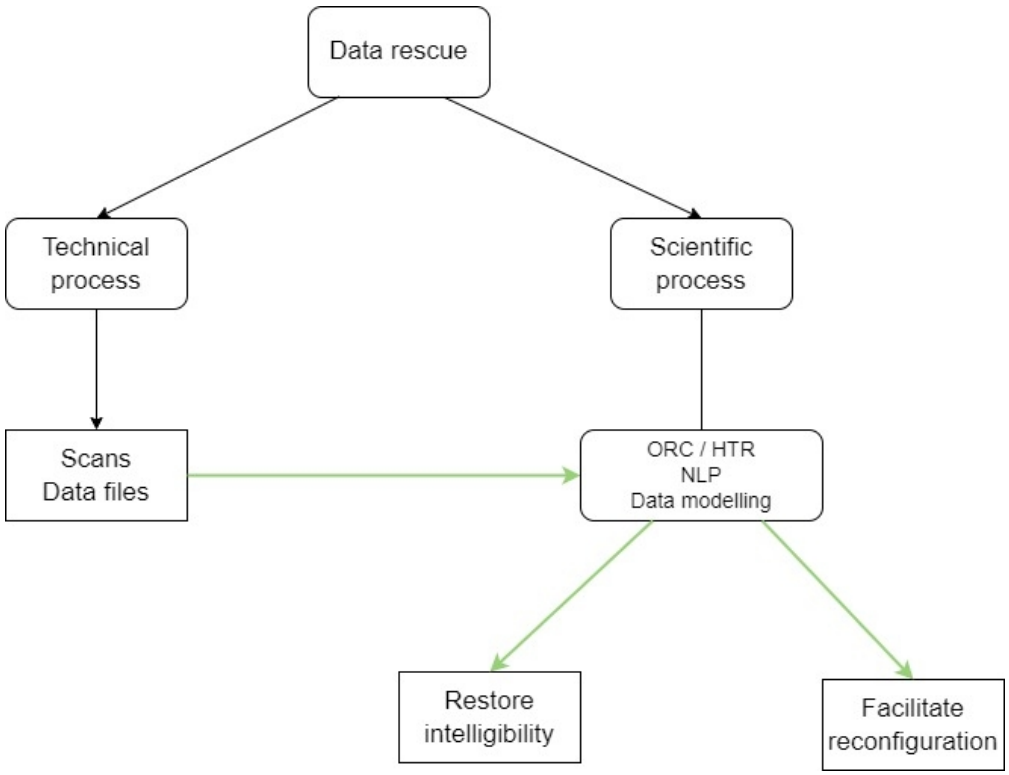
Wilkinson, M., *et al.* (2016), "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3. https://doi.org/10.1038/sdata.2016.18

Zimring, C.A. (2016), "Upcycling in History: Is the Past a Prologue to a Zero-Waste Future? The Case of Aluminum", *RCC Perspectives*, Vol. 3, pp.45-52. http://www.jstor.org/stable/26241375

Zupko, R.E. (1985), *A Dictionary of Weights and Measures for the British Isles: The Middle Ages to the twentieth century*. American Phisolophical Society, Philadelphia.

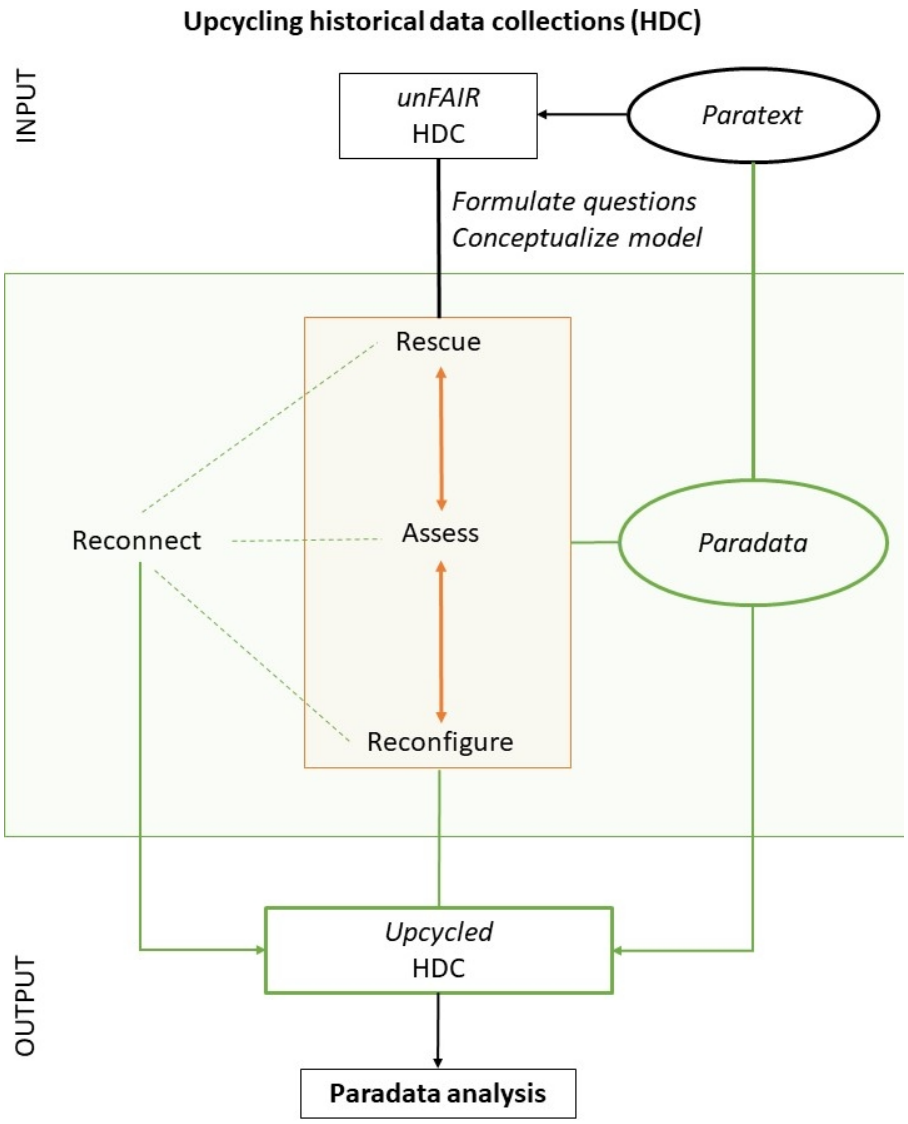
	Historical metrology	*Maritime history*
Pre-1950 HDCs	Noback & Noback 1851	Sound Toll Tables
Post-1950 HDCs	Zupko 1985	Liverpool Shipping and Trade
	Reference works	*Legacy collections*

Table 1: Overview of example HDCs in the text.



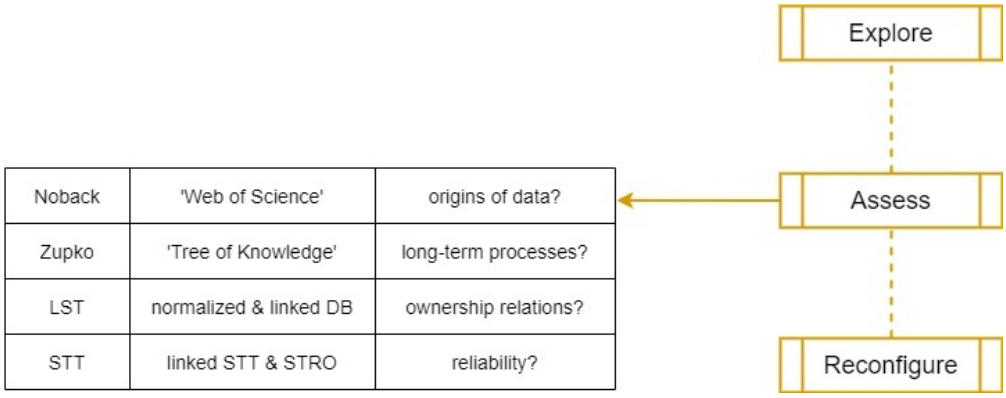
Value gains from data rescue

233x178mm (72 x 72 DPI)



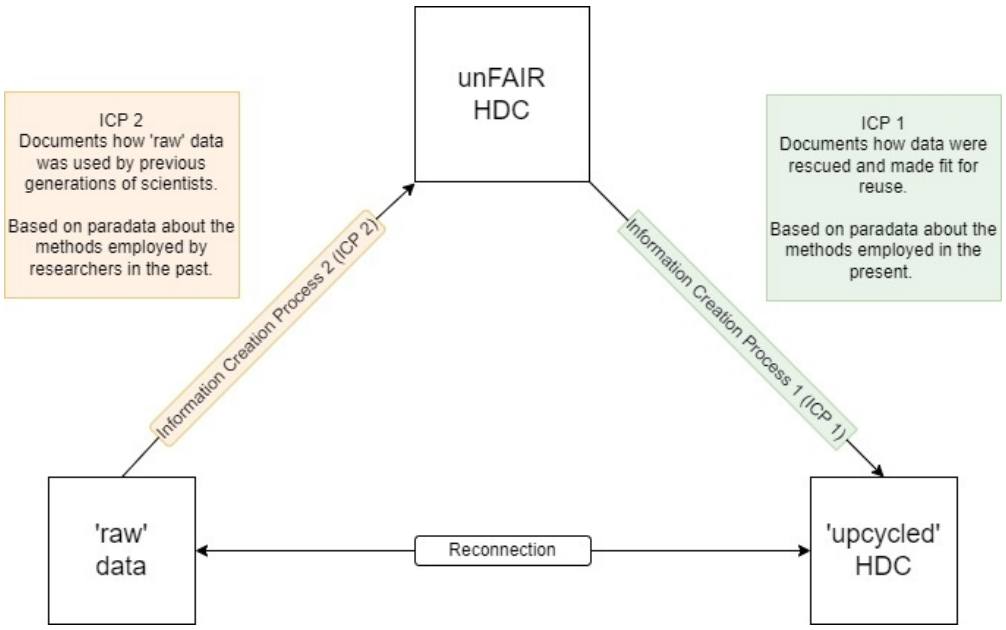
The upcycling workflow

190x254mm (96 x 96 DPI)



Applications of the scenario-based method

254x99mm (72 x 72 DPI)



The study of information creation processes using paradata analysis

240x148mm (72 x 72 DPI)