Universiteit
Leiden
The Netherlands

Knowledge extraction from archives of natural history collections

Stork, L.

Citation
Stork, L. (2021, July 1). *Knowledge extraction from archives of natural history collections*. Retrieved from https://hdl.handle.net/1887/3192382

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3192382

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/3192382 holds various files of this Leiden University dissertation.

Author: Stork, L.
Title: Knowledge extraction from archives of natural history collections
Issue date: 2021-07-01

CHAPTER 3

Manuscripts to Databases

"You who read me—are you certain you understand my language?"

– Jorge Luis Borges, *The Library of Babel*

Searching through historical manuscript collections can seem like an insurmountable task. Misreading one word can change the entire reading of a text, and even a correct reading of a historical text might not give any direct clues as to its meaning, as historical content needs to be understood in the spirit of its own time. Tying images of handwritten text to their symbolic representation (such as digital text), allows for computational exploration of the content and facilitates their correct interpretation.

In this chapter, we aim to answer research question **Q.1** (*What are the trade-offs of various system designs for the disclosure of digital archives?*).

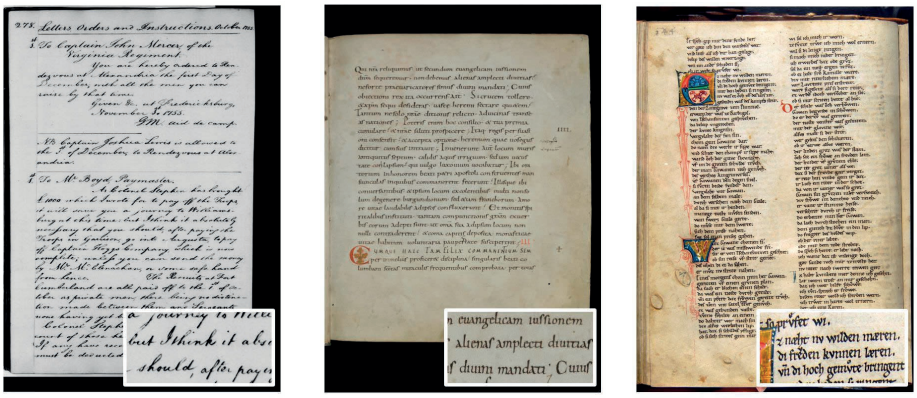
3.1 Introduction

Galleries, Libraries, Archives and Museums (GLAMs) often provide web-accessible, digitised images of historical manuscripts from various domains, e.g., medieval manuscripts,[1] natural history field books,[2] works on philosophy and jurisprudence,[3] ancient religious manuscripts,[4] notarial acts,[5] or biodiversity literature.[6]

In order to computationally access the content of text images, they can be transcribed and/or annotated by the public at large through crowdsourcing (61; 17), or by human domain experts through nichesourcing (62; 63). By utilising human-generated transcriptions,

[1] https://dlmm.library.jhu.edu/en/digital-library-of-medieval-manuscripts/
[2] https://siarchives.si.edu/about/field-book-project
[3] https://blogs.ucl.ac.uk/transcribe-bentham/
[4] https://www.deadseascrolls.org.il/about-the-project/the-digital-library
[5] https://alleamsterdamseakten.nl/
[6] https://www.biodiversitylibrary.org/

automated techniques such as HTR (23; 64) and keyword spotting (65) can further take up transcription. Figure 3.1 shows example historical manuscripts from three different datasets available from the comprehensive IAM-HistDB[1] research database:



(a) (b) (c)

Figure 3.1: Sample pages from the: (a) George Washington, (b) Saint Gall, and (c) Parzival datasets, taken from the IAM-HistDB research database.

Computational systems that produce machine-readable content from historical manuscripts, such as the ones in Figure 3.1 commonly contain three components that each digest the output from the previous component (see also Figure 3.2) (66):

Comp.1 Pre-processing of the heterogeneous content through *document image analysis* (DIA): e.g., segmentation of the heterogeneous content into page elements such as paragraphs, lines and word zones.

Comp.2 Manual or automated transcription of the segmented lines or word zones.

Comp.3 Some form of information extraction or retrieval techniques. The former often by means of natural language processing (NLP) techniques over transcribed texts.

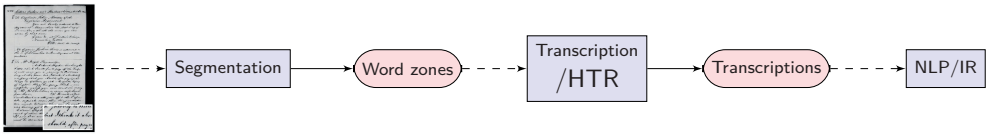


Figure 3.2: Three typical steps in historical document processing. Blue square boxes indicate processes while red rounded boxes indicate output of these processes.

[1] https://diuf.unifr.ch/main/hisdoc/iam-histdb

In this chapter we discuss various systems used in the literature for the enrichment of historical manuscripts. We divide the systems into three groups based on a set of properties that we define (Section 3.2). Based on a final discussion, we propose an approach for knowledge extraction from digital images of field books and scientific illustrations (Section 3.3).

3.2 System Designs

We analyse systems for the enrichment of manuscripts in a slightly less conventional way, for the purpose of optimising and streamlining knowledge extraction. In the literature, systems are often discussed based on types of algorithms used for **Comp.1** (related to binarisation, segmentation, text-line normalisation (66)) and for **Comp.2**, techniques for HTR and OCR and their performance on standard benchmarks (such as the ones from Figure 3.1). **Comp.3** is often looked at separately, after realisation of **Comp.1** and **Comp.2**. We focus on component **Comp.2** and **Comp.3** in conjunction, and look at three properties in specific: **agents** that aid in the transcription and annotation process, the **proportion** of the content that is transcribed, and **richness** of content descriptions:

- **Agents:** The agents that are involved in the process of digitisation of the text: (1) the public at large, (2) the expert community, (3) a machine.
- **Proportion:** The proportion of text that is transcribed, whether it is attempting full verbatim transcription or retrieval of keywords, in which each step includes the previous step(s): (1) named entities, (2) keywords (3) full text.
- **Richness:** The level of richness with which the content is described, in which each step assumes employment of the previous step(s): (1) verbatim, (2) locally defined semantic tags, (3) terms from controlled vocabulary or schemas, (4) IRIs, (5) terms from an ontology.

We define a set of terms in the context of manuscript enrichment, as the terminology may vary between studies:

- **Transcription:** The digital representation of a written text. *Transcribing* in this context is the act of transforming the verbatim handwritten text in a digital image of a manuscript to digital text.
- **Label:** The representation of a region of interest (ROI) in a digital image as digital text. *Labelling* in this context is the act of "attaching" a digital label to a ROI using some computational system. The ROI together with its representation as digital text can be used as training data for machine learning.

- **Annotation:** digital or written notes or comments added to an image or digital text; they point to a specific ROI (for images) or range (for digital text), and add comments or metadata such as a free text description or a semantic type (*semantic annotation*).
- **Keyword:** A word that is key in describing the content of a document, such as a word that would be used to search a set of documents using a search engine.
- **Named entity**: "Information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions (30)."

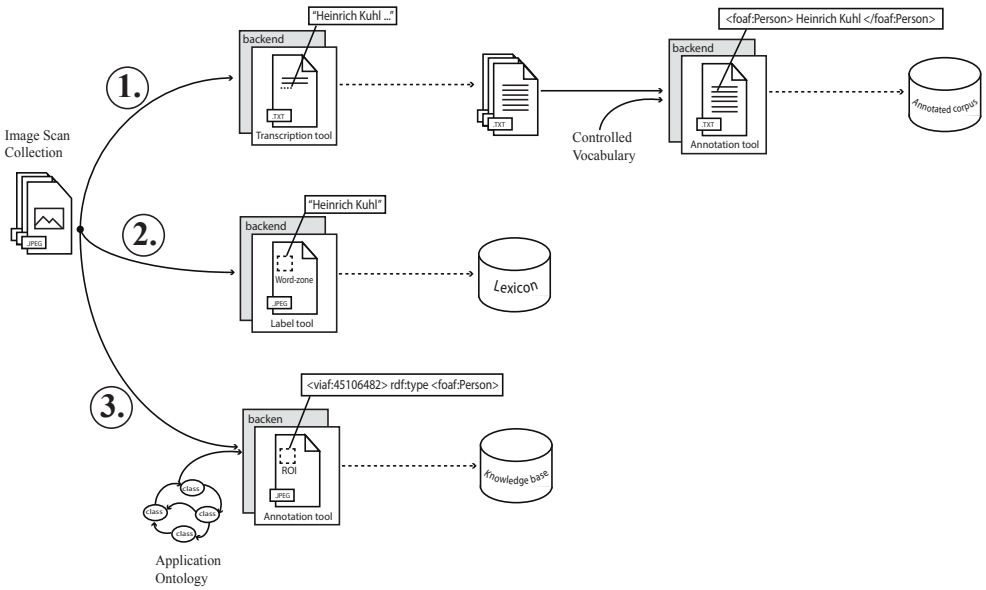


Figure 3.3: Manuscript enrichment design patterns.

We discuss our predefined properties for three types of systems: (1) manual full-text transcription, (2) semi-automated transcription, and (3) semantic annotation of text images, which we graphically represent in Figure 3.3. In total, we discuss a selection of 10 systems that in our opinion represents the breadth of the literature well. In the coming subsections we describe each system type, for which we discuss example frameworks and projects.

3.2.1 Manual Full-Text Transcription.

GLAMs around the world are beginning to notice the potential of crowdsourcing for full-text transcription (see Figure 3.3, system type (1)). In crowd- or nichesourcing, scholars,

experts or the public at large, take on the task of digitisation of the verbatim content of manuscripts (17; 61; 62). Many examples of initiatives now exist that manually transcribe manuscripts in full. We discuss three examples below, specifically including two that digitise handwritten field books:

- The *Field Book Project*[1] (15) is a project set up by the Smithsonian Institution Archives in collaboration with the National Museum of Natural History. The project uses a crowd of what they call "volunpeers"[2] to harvest full-text transcriptions from field books (67), through their transcription center.[3] Controlled vocabularies such as the Natural Collections Description (NCD) are used to describe metadata on levels above content-level (see Figure 1.1). Their approach, described in (68), mentions the use of geo-tagging for future work, to disambiguate localities.

- The *Transcribe Bentham* initiative[4] has digitised and, through crowdsourcing, successfully transcribed 24,833 (update: 27th of November 2020) manuscript pages from jurist Jeremy Bentham (1748-1832), stored in the University College London digital archive, through a customised version of the MediaWiki[5] transcription interface (17; 61).[6] Manuscripts are transcribed, and transcriptions are marked-up with Text Encoding Initiative (TEI)[7]-compliant XML. They indicate a survey pointed out most volunteers took an interest in the history and life of *Bentham*, and that reasons which kept volunteers from transcribing were difficulties deciphering the hand of Bentham. Within another project, *tranScriptorium*[8] (22), the transcriptions are used to further transcribe the manuscripts using HTR techniques.

- The project *From Documents to Datasets* (35) provides a design for the conversion from digitised handwritten field books to datasets, see Figure 4.2, structured according to terms from the DwC standard. They propose first to fully transcribe the texts together with experts, then upload those texts together with the image scans to a MediaWiki[9] server. Via templates, the *taxa*, *locations* and *dates*, are annotated by researchers through a crowdsourcing initiative. Annotators can resolve verbatim names to current ones (taxonomic referencing) during the semantic annotation process. The

[1] https://siarchives.si.edu/about/field-book-project

[2] A combination of the word *volunteer* and *peer*. The term is coined by Meghan Ferriter of the Smithsonian Transcription Center, and is used to refer to a skilled volunteer working at a professional level. https://siarchives.si.edu/blog/growing-community-volunpeers-communication-discovery

[3] https://transcription.si.edu/

[4] https://blogs.ucl.ac.uk/transcribe-bentham/

[5] https://www.mediawiki.org/wiki/MediaWiki

[6] https://blogs.ucl.ac.uk/transcribe-bentham/

[7] TEI is a standard for the representation of texts in digital form, in order to represent structure and content of the text, such as page layout and physical properties https://tei-c.org/

[8] http://transcriptorium.eu/

[9] https://wikisource.org/

annotations are then extracted and converted manually to DwC terms, in order to publish them to the GBIF [1] data server (69).

Agents. Full-text transcription offers a good solution for GLAMs aiming to digitise their manuscript collections, but we note that manuscripts with heterogeneous hard-to-read historical handwriting and content can be too challenging to transcribe by the public at large (**Chall.6**, in line with **Chall.7**). Multiple crowdsourcing techniques exist that secure data quality,[2] but motivation can drop when tasks are too challenging. Although transcription projects often mention they leverage the crowd, most valuable effort appears to come from the community (domain enthusiasts, volunpeers, domain experts, citizen scientists, amateur experts). Transcription and annotation of heterogeneous, multilingual, hard-to-read manuscripts is a knowledge-intensive task, and (amateur) experts have more domain knowledge to perform the tasks, and are intrinsically motivated to produce high-quality data (62). In this sense we note that the term crowdsourcing is an ambiguous one, as there is a significant distinction between the public at large, and the smaller community crowd. We therefore prefer to use the term *nichesourcing* (coined by de Boer et al. (62)) to refer to the act of leveraging a smaller "crowd" of domain (amateur) experts for such knowledge-intensive tasks.

Proportion. As the term *full*-text transcription suggests, the aim of most crowd- or nichesourcing efforts through transcription tools aim at transcribing a text in full. One thing to note is that full-text transcription is time-consuming, and success depends on many factors, such as the complexity of the material and the involvement (motivation) of the community crowd. Full-text transcription can mitigates semantic enrichment, since the manipulation of digital text is computationally more straightforward than the manipulation of text images. However, much of the digitised textual content serves human comprehension, the "glue" that connects the truly interesting pieces of information, and is often not used as search terms.

Richness. Although some systems discussed above employ some form of semantic enrichment (richness level 2), most transcription systems in the literature, however, produce unstructured or semi-structured—usually based on syntax rather than semantics—XML files. This is useful for further searching and processing (e.g. using text mining techniques), but does not enable content to be semantically queried, or integrated with other collections.

[1] http://www.gbif.org/

[2] http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html

3.2.2 Semi-Automated Transcription

Transcription can be partly taken on by HTR techniques (see Figure 3.3, system type (2)). Human experts take on the task of labelling segmented lines or word zones (ROIs containing written words), which are in turn used to automatically increase searchability of other parts of the text. An increase in human-generated transcriptions invokes an increase in the ability of HTR and word spotting techniques to accurately transcribe words in other parts of the texts. Common techniques include supervised deep learning methods such as BLSTMs for classification of characters, full words or sentences, or clustering techniques such as keyword spotting, where "clouds" of visually similar word zones are labelled by experts, rather than single word zones. In our discussion we omit systems that employ OCR, as the content of historical manuscripts is too heterogeneous (see **Chall.6**) for OCR to produce any usable results.

- The *HisDoc* project[1] is an example of a HTR system: experts transcribe individual text lines, and these are used as input to a supervised learning system that aims to learn models for single characters (64). As their system performs HTR at word level, a lexicon (a set of valid words) is required for automated transcription. As an alternative, they experiment with lexicon-free word spotting techniques (65). In the literature, keyword spotting is referred to as a *recognition-free* approach (70): word images are matched to visually similar images, often through a form of clustering of word images in a feature space (71). In order to deal with name variants and misspellings, they define word confusion candidates as synonyms (72).

- A 17th-century botanical manuscript "*Historia de las plantas*" has been digitised (73), using the the Computer Assisted Transcription of Text Images (CATTI) framework (74; 75). The framework performs layout analysis and allows users to transcribe the extracted line segments. The framework also offers HTR technology as an "assistant" that helps users transcribe the text. The HTR technology is based on Hidden Markov Models (HMMs) that operate on single characters, and language models that use as input N-grams. Toselli et al. (73) indicate that the CATTI system primarily aims at producing high-quality professional manuscripts, but indicate that potentially, the crowd could be leveraged, as was done in the *Transcribe Bentham* project.

- *Transkribus* is a platform developed for the enrichment and searching of historical documents (76). A user can transcribe sentences which are then used for training using HTR (21). Similarly to the HisDoc project, Transcribus uses keyword spotting techniques that allow users to search the texts. The project implements a form of

[1] https://diuf.unifr.ch/main/hisdoc/

semantic enrichment: users can use locally defined, user-created semantic tags to label transcriptions or segments.

- The *MONK* system is a search engine for processing multilingual, multi-script historical text, developed by Schomaker (23). It implements HTR as a function for word retrieval. The goal of MONK is therefore not necessarily full-text transcription, but rather to create a searchable index (77). The system has already processed many documents, amongst which the *Dead Sea Scrolls*;[1] Hebrew manuscripts encountered in the Qumran Caves near the Dead Sea.

Agents. Machines, through HTR techniques, can take part in a transcription effort, but have trouble transcribing content that is too heterogeneous (see **Chall.6**), as good results rely on many human-labelled examples. Character-based methods rely on language models and are therefore dependant on a statistical language model or lexicon, whereas an object recognition approach that looks at whole words (such as the one taken by MONK, or word spotting techniques) has to deal with **Chall.8**, as interesting words lie in the long tail of the word distribution. Historical handwriting recognition is far from solved (23), and especially for heterogeneous content, often produces poor results that are difficult to interpret.

Proportion. It appears that, for many HTR systems and their users, the eventual goal is full-text transcription of complete manuscript collections. Other systems aim at creating a searchable index, which does not necessarily require all content to be transcribed. Ultimately, the process is never linear for HTR systems: more transcriptions lead to an increasing number of accurately recognised words. A partly transcribed collection can also be published online as a "living" document of which the proportion of machine-readable content continues to grow.

Richness. The main goal of HTR systems is verbatim transcription (richness level 1), although some allow for semantic enrichment, often no further than richness level 2. It is worthwhile to note that automated tasks such as NERC that further enrich the verbatim content to capture any implicit semantics commonly rely on NLP, a technique that relies on the context of words rather than words in isolation, and therefore depends on the transcription of that context. Although full-text transcription is not required to make a text searchable (not many scholars would be interested to find all instances of the word "*the*" in a collection), we do argue that *undirected* (as in: unguided by formalisms) word-zone

[1] https://www.deadseascrolls.org.il/about-the-project/the-digital-library

labelling or keyword spotting limits or hampers automated extraction of any semantics before manuscripts are fully transcribed.

3.2.3 Semantic Annotation of Text Images.

GLAMs make increasing use of Semantic Web technologies to enrich and publish their collection items (78; 79; 80). Several systems on the web aim for semantic annotation of textual resources (31; 81), but digitised manuscripts are not often enriched in the same way. There are, however, a couple of example systems that directly annotate text images with semantic concepts. Similarly to word-zone labelling, scholars, experts or the public at large can be employed to semantically annotate online documents (see Figure 3.3, system type (4)).

- *Accurator*[1] is an example of a web application that uses an expert crowd to annotate digital images, in specific digitised items from cultural heritage collections, such as paintings. Web users can help museums describe their collection items by providing expert knowledge. They are prompted to annotate digital renditions of items from cultural heritage collections with terms from controlled vocabularies, carefully selected for the target domain of the collection. For each collection, experts were even involved in the process of determining a goal for proper enrichment, in order to improve access to the collection in question. Annotations are stored in RDF format and linked to the digital images using the Web Annotation Vocabulary[2] (82).

- Ebert et al. (2010) (83) perform ontology-based information extraction (OBIE) from handwritten documents. They are one of the first ones to introduce the topic to the field of HTR. Interestingly, their system employs a dialogue between a component that deals with HTR and a OBIE component. Their system is based on digital ink as input (using the MyScript[3] system for HTR) and the scope of their experiments is homogeneous handwriting (they experiment with modern English handwritten texts) rather than the heterogeneous material from historical manuscript collections, which additionally needs to deal with historical multilingual text (**Chall.6**).

- Adak at al. (2016) (84) perform named entity recognition (NER) on unstructured handwritten text images, without employing any character or word recogniser. After word segmentation, they extract engineered structural and positional features from word zones, which are used in a BLSTM for NER. Classification of the named entities is out of the scope of their paper. The methodology presented in the paper does not

[1] http://www.accurator.nl/
[2] https://www.w3.org/TR/annotation-vocab/
[3] https://developer.myscript.com/docs/concepts/introduction/

increase searchability of the text, but can be combined with a controlled vocabulary for NERC to automatically enrich the handwritten content semantically. We therefore included it in this section. The article presents a nice overview of how relevant page elements such as named entities can be identified in text images with hard-to-read historical texts.

Agents Semantic annotation of texts is a more knowledge-intensive task than mere verbatim transcription of a text, as a level of interpretation is required. Therefore, human (amateur) experts are required to take part in the annotation process. Additionally, quite some time is spent selecting or re-engineering vocabularies or ontologies to fit the target domain. However, an application ontology formalises the minimal information required for annotation, thereby driving the enrichment process. Moreover, machines can take part in the semantic annotation process, as is shown by Adak et al. (84).

Proportion The systems mentioned above operate on text (or multimodal) images, and focus on the annotation of information units, such as named entities, rather than just any word or full text. Prior to the annotation effort, the expert community decides on interesting concepts and their meanings, and use these to semantically enrich ROIs through a nichesourcing initiative, which users eventually use to navigate and understand the resulting knowledge base, and join distributed collections.

Richness At a minimum, semantic annotation systems annotate texts or text images with semantic concepts, for instance through a combination of supervised HTR and NERC from features of the handwritten text (85; 84; 86) (richness level 2). Examples exist that even use terms from controlled vocabularies or schemas (richness level 3), or that use HTTP URIs for better content descriptions (richness level 4) (82; 31; 81).

3.3 More Product, Less Process

Coming back to **Q.1** (*What are the trade-offs of various system designs for the disclosure of digital archives?*), we note that the enrichment of manuscripts is often a highly time-consuming process that depends on community engagement. This is no different for field book manuscripts, which are exceptionally challenging to make sense of, given **Chall.1** to **Chall.5**.

At the same time, if we look back at **Chall.6** to **Chall.8**, we note that it seems unavoidable that humans play a large part in the enrichment process, although machines can be employed

to speed up this process, given that their results are presented in a transparent, human-understandable way. Systems with high recall but low precision[1] increase retrievability of words, but results can clutter the enrichment process when not presented well. Moreover, unless character-based out-of-lexicon methods are employed, words that occur more often are the first to be recognised accurately, while they are more likely to be less relevant. A third thing to note is that enrichment efforts often result in unstructured or syntactically structured digital text, that require a crucial enrichment step in order to be understood and reused by scholars and the general public.

We have observed in Subsection 2.1.2 and systems discussed in the previous section, that the content in manuscripts from NHCs is organised around a systematic regularity that is intrinsic to the field of biodiversity, in which researchers attempt to systematise the natural world. This systematic organisation is not commonly encountered in other manuscripts. At the same time, community standards are set up to formalise these systematics.[2] In terms of efficiency; should "volunpeers" not maximise their impact by focussing not only on transcription, but also on systematics, using standard formalisms from the domain?

Greene et al. (87) already noted in 2005 in their article *More Product, Less Process* that there exists a huge backlog of unprocessed archival material (for the most part the authors refer to cataloguing of archives on a collection- and item-level for minimal collection access, but we argue that the same concerns apply to enrichment of and access to archival content). They mention that processing of archival material should: *"describe materials sufficient to promote use."* To strengthen their argument, they quote an article already published three decades ago on the same topic:

> We rarely ask the question: when is *this* collection processed? Instead, we process all collections to an ideal standard level. The second problem is that by processing all collections to the ideal standard level, we cannot keep up with the collections we have on hand or with the new collections coming in. The result tends to be a small number of beautifully processed collections available for use and an extensive backlog of collections that are closed while they wait to be processed (88).

This idea is in line with the idea of Minimum Information about a Digital Specimen (MIDS)[3] from the Collection Descriptions (CD) interest group, on the formalisation of sufficient digitisation:

> A harmonizing framework captured as a TDWG standard can help clarify levels (depth) of digitization and the minimum information captured and published at each level.This would help to ensure that enough data are captured, curated and published against specific requirements so they are useful for the widest range of

[1] recall refers to the percentage of *all* words that is correctly retrieved, while precision refers to the percentage of words that is correctly retrieved from all *retrieved* words.

[2] `https://www.tdwg.org/`

[3] `https://www.tdwg.org/community/cd/mids/`

> possible purposes; as well as making it easier to consistently measure the extent of digitization achieved over time and to set priorities for remaining work (89).

We extend these ideas to the digitisation of manuscript *content*. We claim that at a minimum, information extraction from manuscripts should promote document understanding, rather than full-text transcription of each manuscript to an ideal level.

We therefore opt for a targeted approach, in which the expert community decides the semantic concepts relevant for document understanding and search, maps these to existing ontologies and IRIs, and uses these to guide the annotation effort by semantically annotating and transcribing the relevant word zones in text images through a nichesourcing initiative. Texts are made searchable, pointing users to interesting bits of the text documents, while ground truth is generated for semi-automated semantic annotation (similar to NERC) as well as verbatim transcription. In an end-to-end approach, a named entity recogniser can then benefit from output of the handwriting recogniser, and vice versa.

Although some extra work is required to semantically annotate texts with Linked Data (LD), omitting full-text transcription means having to annotate only a small percentage of the content; e.g., focussing on the transcription and semantic annotation of those *named entities* that allow users to construct rich semantic queries or aggregate informative content across archival collections.

Pre-populating knowledge bases with background knowledge, such as collection-specific locations from the Geonames database or collection-specific persons from the Virtual International Authority File (VIAF) authority IRIs, helps annotators to use the correct named entities for annotation. Using LD for annotation helps remove ambiguity as IRIs contain rich descriptions. The name "*Heinrich Kuhl*", for instance, is ambiguous. If we instead use the IRI `https://viaf.org/viaf/45106482/`, we agree on the reference of the verbatim name to the person "*Heinrich Kuhl*" (1797-1821), a German zoologist.

Lastly we argue that annotation *provenance* is a dimension that is often overlooked, but should be seen as a critical step in the elucidation process. With data provenance we refer to data concerning the lineage of data: why, when, and how they were produced or changed, and measures of their quality (90; 91; 92). Storing provenance of annotations contributes to publishing annotation knowledge graphs in a FAIR way, allowing scholarly discussions over the content and reproducibility of hypotheses and results.