From Handwritten Manuscripts to Linked Data

Lise Stork[1,2(✉)], Andreas Weber[3], Jaap van den Herik[1,2], Aske Plaat[1,2], Fons Verbeek[1,2], and Katherine Wolstencroft[1,2]

[1] Leiden Institute of Advanced Computer Science, Leiden, The Netherlands
{l.stork,k.j.wolstencroft,f.j.verbeek,a.plaat}@liacs.leidenuniv.nl
[2] The Leiden Centre of Data Science, Leiden, The Netherlands
h.j.vandenherik@law.leidenuniv.nl
[3] University of Twente, Enschede, The Netherlands
a.weber@utwente.nl

Abstract. Museums, archives and digital libraries make increasing use of Semantic Web technologies to enrich and publish their collection items. The contents of those items, however, are not often enriched in the same way. Extracting named entities within historical manuscripts and disclosing the relationships between them would facilitate cultural heritage research, but it is a labour-intensive and time-consuming process, particularly for handwritten documents.

It requires either automated handwriting recognition techniques, or manual annotation by domain experts before the content can be semantically structured. Different workflows have been proposed to address this problem, involving full-text transcription and named entity extraction, with results ranging from unstructured files to semantically annotated knowledge bases. Here, we detail these workflows and describe the approach we have taken to disclose historical biodiversity data, which enables the direct labelling and semantic annotation of document images in hand-written archives.

Keywords: Linked data · Cultural heritage
Handwriting recognition · Semantic annotation
Named entity recognition

1 Introduction

Digital libraries often provide web-accessible, digitised images of handwritten manuscripts from various domains. However, the challenge remains to elucidate the handwritten content in a way that will enable exploration and further research. This involves the transformation of the content into a searchable knowledge base. Historical documents are especially difficult due to the hard-to-read handwriting, often in multiple languages, and the historical context of the text, which makes them difficult to interpret. To enrich and elucidate the content of manuscripts, different workflow methods have been developed. Digitised images

E. Méndez et al. (Eds.): TPDL 2018, LNCS 11057, pp. 330–334, 2018.
https://doi.org/10.1007/978-3-030-00066-0_34

of content can be annotated by human domain experts through nichesourcing [3], or computationally using automated handwriting recognition or word spotting techniques [2,7]. Most workflows, however, produce flat files or semi-structured output. This is useful for further searching and processing (e.g. using text mining techniques), but it does not enable content to be interlinked, semantically queried, or compared to other collections. We argue that this can be facilitated by labelling *and* semantically annotating word-zones - single word segments extracted from document images - using a domain ontology, resulting in a rich knowledge base that can be queried and interlinked with external resources.

2 Workflows for Elucidating Contents

Figure 1 roughly presents common workflows for the enrichment of handwritten documents.

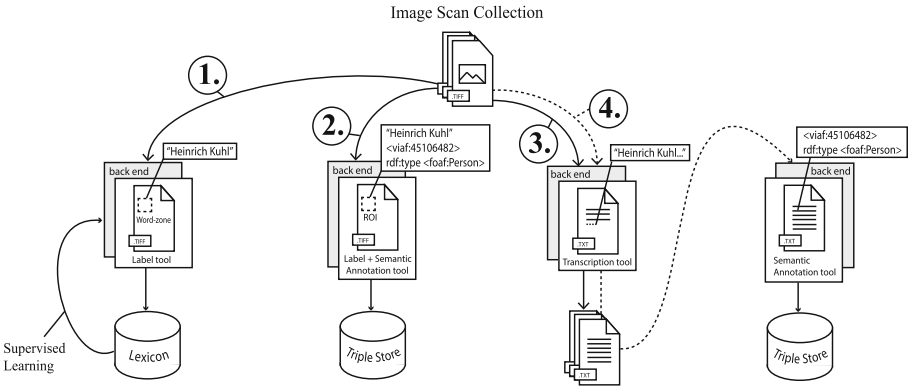

Fig. 1. Manuscript enrichment workflows

Workflow 1. The Automated Labelling of Word-Zones in Images. The *HistDoc* project is an example of a *Handwritten Text Recognition* (HTR) system that uses experts to harvest labels as input to a learning system [2]. Another example is *Transkribus*, where users can label sentences which are then used for training using HTR [5]. This project implements a form of semantic enrichment: labellers can flag certain named entities, e.g., locations or persons, with a user created tag set. Lastly, the aim of the *MONK* handwriting recognition system [7] is not full-transcription per se, but rather searchability of the informative content. The system does not rely on a language model and is therefore adaptive to its input. Labelling is targeted: the system retrieves and labels words that are visually similar to word-zones that are labelled by users.

Workflow 2. Semantic Annotation of Manuscript Images. *Accurator*[1] 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. Users are prompted to annotate cultural heritage items with carefully selected controlled vocabularies. Annotations are stored in RDF format and linked to the digital images using the Web Annotation Data Model [3]. Another example, the *Semantic Field Book (SFB) Annotator*[2] [8], labels and simultaneously semantically annotates the most informative content of digitised manuscripts from natural history collections using an application ontology and the Web Annotation Data Model.

Workflow 3. Full-Text Transcription. The *Field Book Project*, a collaboration between the Smithsonian Institution Archives and the National Museum of Natural History, uses the crowd to harvest full-text transcriptions from historical biodiversity field books [1]. Another example is the *Transcribe Betham* initiative that will digitise and, also via crowdsourcing, fully transcribe 12,500 folios from the jurist Jeremy Bentham (1748–1832), stored in the University College London digital archive, through a media-wiki interface [6].

Workflow 4. Semantic Annotation of Fully Transcribed Text. *Annotea*[3] is a shared web annotation system which is based on the semantic annotation of web-based text files. In the Annotea architecture, annotations exist externally from the documents on *annotation servers*. The system lets an annotation point to a piece of digital text using the XPointer framework.[4] Other users are able to add their own additional annotations. Annotea makes use of existing W3C specifications, such as RDF and HTTP [4]. Another example is the *From Documents To Datasets* project [9]. Biodiversity field books are first fully transcribed and then semantically enriched.

Combining Automated Word-Zone Labelling with Semantic Annotation. In the Making Sense project,[5] methods are being developed for automated semantic annotation of natural history collections [10]. Our use case consists of 8,000 field book pages gathered by the Committee for Natural History of the Netherlands Indies between 1820 and 1850. A field book contains records that report species observations: their anatomy, characteristics, habitat and behaviour. Aiming for targeted, semantic annotation, the Making Sense project currently operates workflow 1, through the *MONK* handwriting recognition system, and workflow 2, through the *SFB-Annotator*. Initial results, an ontology

[1] http://www.accurator.nl/.
[2] https://github.com/lisestork/SFB-Annotator/.
[3] https://www.w3.org/2001/Annotea/.
[4] https://www.w3.org/TR/xptr-framework/.
[5] http://www.makingsenseproject.org/.

and a web application (see footnote 2) are available. Our final goal, however, is to combine workflows 1 and 2. Expert curated labels and semantic annotations can be used as input to a supervised learning system, combining handwriting and named entity recognition to perform semi-automated semantic annotation, thereby streamlining the process of elucidating, labelling and interlinking named entities.

3 Conclusion and Future Work

In this study, we enumerated the different workflow approaches that have been used to extract and structure the content of historical manuscripts, illustrated by example projects that utilise them. Although full-text transcription is an effective procedure that is often used, it cannot scale for all archived data and it falls short for further exploration and interpretation. Tools should be developed that reduce the requirement for full-text transcription and facilitate semantic annotation of extracted text to enable richer content descriptions. In our case study, we show that by providing tools to enable the direct semantic annotation of named entities, we can reduce the full-text transcription burden. In future work we will develop automated methods for semantic annotation.

By publishing the results online as Linked Open Data, the contents can be disclosed as a rich, structured resource that can be searched and combined with other cultural heritage collections.

References

1. The Field Book Project. https://siarchives.si.edu/about/field-book-project. Accessed 14 Mar 2018
2. Baechler, M., Fischer, A., Naji, N., Ingold, R., Bunke, H., Savoy, J.: HisDoc: historical document analysis, recognition, and retrieval. In: Proceedings of Digital Humanities, pp. 94–96. University of Hamburg, July 2012
3. Dijkshoorn, C., De Boer, V., Aroyo, L., Schreiber, G.: Accurator: nichesourcing for cultural heritage. Computing Research Repository, abs/1709.09249 (2017)
4. Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., Swick, R.R.: Annotea: an open RDF infrastructure for shared web annotations. Comput. Netw. **39**(5), 589–608 (2002)
5. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 4, pp. 19–24. IEEE (2017)
6. Moyle, M., Tonra, J., Wallace, V.: Manuscript transcription by crowdsourcing: transcribe bentham. Liber Q. **20**(3–4), 347–356 (2011)
7. Schomaker, L.: Design considerations for a large-scale image-based text search engine in historical manuscript collections. IT - Inf. Technol. **58**(2), 80–88 (2016)
8. Stork, L., et al.: Semantic annotation of natural history collections. Web Semant.: Sci. Serv. Agents World Wide Web. (2018). https://doi.org/10.1016/j.websem.2018.06.002

9. Thomer, A., Vaidya, G., Guralnick, R., Bloom, D., Russell, L.: From documents to datasets: a mediawiki-based metod of annotating and extracting species observations in century-old field notebooks. ZooKeys **209**, 235–253 (2012)
10. Weber, A., Ameryan, M., Wolstencroft, K., Stork, L., Heerlien, M., Schomaker, L.: Towards a digital infrastructure for illustrated handwritten archives. In: Ioannides, M. (ed.) ITN-DCH 2017. LNCS, vol. 10605, pp. 155–166. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75826-8_13