Towards a digital infrastructure for illustrated handwritten archives

Andreas Weber, Mahya Ameryan, Katherine Wolstencroft, Lise Stork, Maarten Heerlien, and Lambert Schomake

Towards a Digital Infrastructure for Illustrated Handwritten Archives

Andreas Weber[1(✉)], Mahya Ameryan[2], Katherine Wolstencroft[3], Lise Stork[3], Maarten Heerlien[4], and Lambert Schomaker[2]

[1] BMS-STePS, University of Twente, 7500 AE Enschede, The Netherlands
a.weber@utwente.nl
[2] ALICE, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands
{m.ameryan,l.r.b.schomaker}@rug.nl
[3] LIACS, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
k.j.wolstencroft@liacs.leidenuniv.nl
[4] Naturalis Biodiversity Center, PO Box 9517, 2300 RA Leiden, The Netherlands
maarten@heerlien.net

Abstract. Large and important parts of cultural heritage are stored in archives that are difficult to access, even after digitization. Documents and notes are written in hard-to-read historical handwriting and are often interspersed with illustrations. Such collections are weakly structured and largely inaccessible to a wider public and scholars. Traditionally, humanities researchers treat text and images separately. This separation extends to traditional handwriting recognition systems. Many of them use a segmentation free OCR approach which only allows the resolution of homogenous manuscripts in terms of layout, style and linguistic content. This is in contrast to our infrastructure which aims to resolve heterogeneous handwritten manuscript pages in which different scripts and images are narrowly intertwined. Authors in our use case, a 17,000 page account of exploration of the Indonesian Archipelago between 1820–1850 ("Natuurkundige Commissie voor Nederlands-Indië") tried to follow a semantic way to record their knowledge and observations, however, this discipline does not exist in the handwriting script. The use of different languages, such as German, Latin, Dutch, Malay, Greek, and French makes interpretation more challenging. Our infrastructure takes the state-of-the-art word retrieval system MONK as starting point. Owing to its visual approach, MONK can handle the diversity of material we encounter in our use case and many other historical collections: text, drawings and images. By combining text and image recognition, we significantly transcend beyond the state-of-the art, and provide meaningful additions to integrated manuscript recognition. This paper describes the infrastructure and presents early results.

Keywords: Deep learning · Digital heritage · Natural history
Biodiversity heritage

Mahya Ameryan and Andreas Weber share the first authorship of this paper.

M. Ioannides (Ed.): ITN-DCH 2017, LNCS 10605, pp. 155–166, 2018.
https://doi.org/10.1007/978-3-319-75826-8_13

1 Introduction

Many heritage collections and archives consist of documents in which handwritten text in different languages and scripts are interspersed with images. In order to open up and interlink such multimedial collections, heritage institutions typically resort to manual enrichment methods such as keyword tagging and full-text transcription [1–3]. Often, these methods rely on crowdsourcing, where volunteers take large parts of the work upon themselves [4–6]. Although such practices produce high-quality data, it is a labour-intensive, time consuming and therefore costly way of opening up heterogeneous collections [7]. Furthermore, these methods require an advanced level of expertise from professionals and even from volunteers. One can neither transcribe illustrated handwritten manuscripts without thorough knowledge of palaeography, a dying expertise, nor can one add useful subject information to a document or drawing without having knowledge about its context and semantic structure [8]. The disclosure of scientific manuscript collections, for instance, heavily depends on the availability of domain-specific background knowledge [9]. Multimedial manuscript collections cannot be enriched, if one is unable to situate notes and hand-drawn sketches and drawings in their historical context.

Traditionally, humanities researchers treat text and images separately. This separation extends to traditional handwriting recognition systems [10–12]. Many of them use a segmentation free OCR approach which only allows the resolution of homogenous manuscripts in terms of layout, style and linguistic content [13, 14]. This is in contrast to our infrastructure which aims to resolve heterogeneous handwritten manuscript pages in which different scripts and images are narrowly intertwined. Authors in our use case, a 17,000 page account of exploration of the Indonesian Archipelago between 1820–1850 ("Natuurkundige Commissie voor Nederlands-Indië) tried to follow a semantic way to record their knowledge and observations, however, this discipline does not exist in the handwriting script. The use of different languages, such as German, Latin, Dutch, Malay, Greek, and French makes interpretation more challenging. On many pages, handwritten text is also intermixed with sketches and drawings. Owing to this complexity, the paper heritage of the Natuurkundige Commissie (further referred to by the acronym NC) has remained largely inaccessible to scholars and the general public.

Since the reliable semantic interpretation of illustrated handwritten heritage collections requires an integrated approach to text and image recognition, this paper describes the basic layout of a user-centred infrastructure which is developed in the context of the research project Making Sense of Illustrated Handwritten Archives (2016–2020). By integrating text and image interpretation, we aim at providing meaningful additions to integrated manuscript recognition. To address this challenge, our infrastructure takes the state-of-the-art word retrieval system MONK as starting point and augments it with page layout and image analysis, and semantic integration. Owing to its visual approach, MONK can handle the diversity of material we encounter in our use case and many other historical collections: text, drawings and images [15]. This paper also entails initial results on the active learning performance of MONK in the context of the NC collection. Combining image and textual recognition into one digital infrastructure, allows for an integrated study of underexplored heritage collections and archives in general. In our

opinion, this is the most promising way to achieve the required level of accuracy for handwritten illustrated collections. Our work is financed by the Netherlands Organization for Scientific Research (NWO) and the Dutch publishing house Brill.

2 Use Case: The Archive of the Committee for Natural History of the Netherlands Indies (1820–1850)

In order to realize the digital infrastructure, we utilize the notes and illustrations of the NC archive as its use case. This extensive corpus, which was composed by 17 naturalists and draftsmen contains a rich account of scientific exploration in the Indonesian Archipelago in the period 1820–1850 (see Fig. 1). The NC charted the natural and economic state of the Indonesian Archipelago, in the nineteenth century a Dutch colony, and returned a wealth of scientific data and specimens which are stored in the archive and depot of the Naturalis Biodiversity Center in Leiden [16, 17]. In addition to the thousands of handwritten notes and drawings, the collection comprises tens of thousands of biological and geological specimens and a four-volume publication on the commission's findings [18, 19]. Owing to its high scientific and cultural value, Naturalis restored and digitized the NC's paper legacy and specimens between 2007 and 2015 with funds from the Metamorfoze and the FCD programme[1].

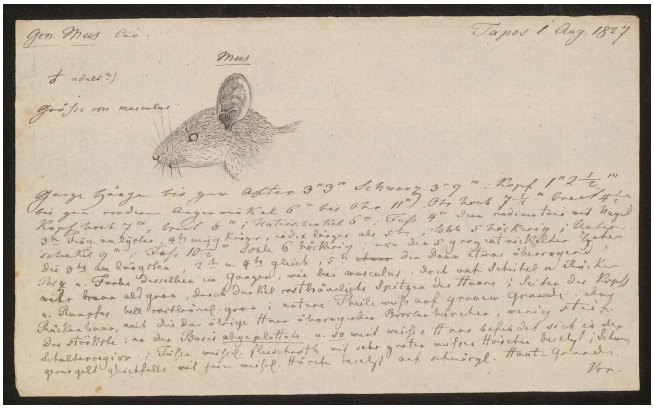


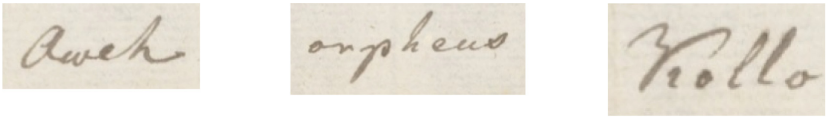
Fig. 1. Illustrated field note by H. Boie (1797–1827). It is composed in a mix of German, Latin, Dutch, Greek, and Malay. Naturalis BC, NNM001001061_020. Public Domain Mark 1.0.

Though the paper heritage of the NC collection is digitally preserved, it remains inaccessible to scholars and the general public, also due to its heterogeneous structure as discussed in the introduction of this paper. In order to establish links between handwritten field notes, drawings and specimens, our digital infrastructure must be able to

[1] The Metamorfoze programme funds the preservation of paper heritage that is deemed to be of national importance for the Netherlands. The FCD programme (FES Collection Digitization, 2010–2015) digitized a significant part of the specimens preserved by Naturalis.

cope with a number of challenges. The biggest challenge on a semantic level is the evolution of concepts; in the case of the NC the evolution of toponyms and taxon names in particular. Since locality names and scientific species names have often changed over time, due to conceptual, political and linguistic shifts, our infrastructure must also be able to deal with background knowledge in the form of controlled vocabularies (e.g. biological taxonomies, gazetteers) and context information as it is provided by publications on individual naturalists [20, 21] and databases [22]. The labels of historic specimens only provide general information on collection localities and collectors (e.g. "Java, Boie"). Until now, only initial attempts were made to disclose and connect the material manually [23].

While authors in our use case tried to follow a structured way to record their knowledge and their observations, however, this discipline does not exist in the handwriting script. Figure 2 show three samples of fully connected, mixed cursive and isolated handprint on a same line of our data. Also, Fig. 3 shows some different binarized labelled word images of NC manuscript showing common problems. Moreover, different languages used in fieldnotes, including German, Latin, Dutch, Greek, Malay and French makes interpretation more challenging. The multi-layered character of the NC collection makes it thus the perfect use case to realize a technologically advanced and usability-engineered digital infrastructure for interpreting illustrated handwritten archives in general.



a b c

Fig. 2. Multiple writing styles on the same line of a NC manuscript: (a), (b) and (c) show fully connected, mixed cursive and isolated handprint styles exist on the same line.

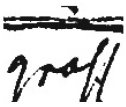
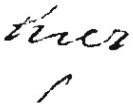
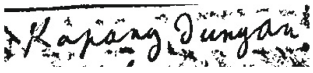
			
Farbe	gross	gross	hier
			
Kapangdungan	nigro	nigra	Theil

Fig. 3. Binarized labelled word images of NC manuscript showing common problems.

2.1 MONK: A Solid Point of Departure

The infrastructure proposed here, takes the state-of-the-art word retrieval system MONK, as a starting point and augments it with page layout, image analysis and semantic integration. MONK achieves a high accuracy on a wide range of script styles [24–26]. It is used by humanities researchers from well-known institutes around the globe (e.g. National Archive and National Library in The Hague, Czech National Archive, Harvard University Yenchin (Chinese handwritten) collection, the Dead Sea Scrolls, in cooperation with Israel Antiquity Authority). MONK was developed in SCRATCH (SCRipt Analysis Tools for the Cultural Heritage), a project in the context of the NWO-funded CATCH programme (Continuous Access to Cultural Heritage, 2004–2014), and its follow-up valorisation project SCRATCH4ALL. The system was scaled up to the 'Big Data' level in TARGET, a project funded by the European Regional Development Fund. Because of its visual approach, MONK can handle the diversity of material that one encounters in many historical collections: text, drawings, and images [15]. The large majority of related systems are designed to handle only a single homogenous manuscript in terms of layout, style and linguistic content.[2]

Internationally, the most noteworthy project aimed at automated historical handwriting recognition is the READ project, a high profile EU e-Infrastructure project funded by the European Commission that is aimed at resolving a small number specific documents in toto [14, 27]. This is in contrast to our goal, where we aim to recognize and semantically interpret pages in a large document collection with a wide variety of writing styles. Since the tool which the READ consortium (Transkribus) is developing needs a single language model, it cannot process different languages and scripts on one page. With our approach, different languages and scripts may be intermixed with images, as the active learning system of MONK does not require prior training and can cope with heterogeneous pages. In our opinion, the tabula-rasa approach of MONK is much better suited for our use case and other historical collections. MONK differs from Transkribus and other systems which follow a segmentation free OCR approach in three ways.

1. The MONK system uses shape-based feature vector methods that have very few assumptions concerning the content or style of the material. The range of different handwriting types, languages and styles that MONK can handle is consequently much broader than that of other handwriting recognition systems. Examples of manuscripts that have been successfully processed by MONK are medieval Western texts, 18th century captain's logs, the Qumran Dead Sea scrolls, Arabic, and Chinese texts. MONK was able to adapt to process Chinese text from scratch, with zero labels within a two-week period [28]. The system currently handles 25 k character classes, processing three hundred manuscripts in wood-block printed and handwritten styles.
2. The reason for this success is the avoidance of the traditional OCR (optical character recognition) approach which assumes that individual characters are essentially legible [15]. This assumption only holds for a fraction of handwritten material and a limited number of scripts. Even where individual characters are seen, by human

[2] An exception is: https://kogs-www.informatik.uni-hamburg.de/projekte/IMPACT.html, last accessed 2017/09/02.

visual inspection, the poor image quality usually requires the full-word pattern to be used for reducing the number of alternative classifications. Several traditional systems, make use of Hidden Markov modelling [29]. This technique has a number of fundamental and practical disadvantages for handling highly diverse image data. We mention three of them. (1) Whilst Markov models work successfully in the one dimensional domain of speech, the two dimensional pixel domain is in many ways much more complex. (2) The amount of labelled data is extremely limited in historical handwriting compared to speech, and (3) hidden-Markov requires model design by highly skilled researchers. This leads to handwriting recognition systems that are manually tuned to a particular application [25]. Live training of such systems by ordinary users is not possible.

3. In the MONK system, active-learning methods are used to allow the system to learn quickly from the input of experts. The system presents a number of word-zone images to the user, who has to confirm their text label, or provide a meaningful label for them. As the system can decide which images to show to an expert, it can collect feedback for those images from which it can learn the most. By combining active learning with learning methods that do not require a large number of labelled examples, such as nearest-neighbour methods, MONK can learn to recognize novel document collections in a relatively short time. As the number of harvested labels increases, increasingly advanced and more 'label-greedy' Machine-Learning methods are employed, such as SVMs, Convolutional Neural Networks, and Multi-dimensional Long Short-Term Memory (MDLSTM) [30]. This allows MONK to keep improving its performance once more examples become available. Alternative methods suggest that by labelling the first few hundred pages of a large collection, you can model the data and you can process the rest of the collection. If the collection is 10,000 or 50,000 pages, however, the initial pages, which are often neatly written, may not be representative of the rest of the collection. A random sampling across the dataset is more suitable for training the model. For this paper, we used MDLSTM for word recognition [31].

The experimental dataset contains 6286 word images with 121 word classes with an alphabet of 37 different characters. Each word class contains 19 to 259 examples. To generalize our results to be independent of the chosen train and test datasets, we used 10-fold cross validation. In order to increase the variation of the training set, they were morphed, two times [32]. For each fold, the original, morphed and two-time morphed images were used for the training set, whilst each of the validation and also test sets contained one of 10 original subsets. We used three configurations for MDLSTM (Table 1). The input block is 2×10, the hidden block size is 3×4 and 2×4; the learning rate is e^{-4}; momentum is 0.9. These architectures differs in hidden and subsample sizes (Table 1). Training was stopped after 30 consecutive evaluations without improvement on the validation set. In order to ensure lexically valid responses, the output strings are corrected by word matching using the Levenshtein distance. The lexical entry with the lowest distance is considered to be the final output. Apart from looking at the most likely hypothesis (top 1), we also counted the number of times a correct word was found among the top-5 candidates. Table 1 shows the mean and standard of deviation of word recognition rate (%) derived by MDLSTM for 10-fold validation on our dataset. It should be

noted that tests such as these are subject to a large number of factors and selection criteria: for classes with many labels, the results will be high, while difficult words usually also have fewer labels, and lower accuracy.

Table 1. Mean and standard of deviation of word recognition rate (%) for three different MDLSTM neural networks, using 10-fold validation.

Arch.	Hidden size	Subsample size	Training	Test (top1)	Test (top5)
A14	8, 40, 80	8, 50	100 ± 00	70 ± 3	83 ± 2
A28	20, 50, 100	8, 4	100 ± 00	73 ± 2	85 ± 1
A30	2, 10, 50	6, 20	92 ± 8	61 ± 2	77 ± 2

2.2 Towards a Digital Infrastructure: Next Steps (2017–2020)

The success of our infrastructure crucially depends on the accurate recognition as well as semantically structuring and interlinking content from handwritten illustrated documents. Our use case is representative of many other cultural heritage collections. In order to achieve this goal MONK, the pattern-recognition basis of which is already solid, is enriched with three new systems which will be discussed separately below.

System I: Layout Analysis System. A layout-analysis system is required that improves the ability to identify and segment visual and textual elements in an image. This entails the inclusion of new, adapted methods for layout analysis, pattern recognition and machine learning. The system ensures that elements in the layout of digitized images are detected and segmented. In order to carry out this task, the system relies on a locally operating segmentation procedure [33] which will be tuned to this type of document data in combination with smart pre-processing to avoid effects of the "background" of the document affecting the quality of the segmentation.

Of special interest in the NC document collection is that many documents have a well-defined, consistent layout; on different pages, images and text may appear at the same place. Moreover, there is often a separation between pictorial material and text in the form of a 'white', i.e. paper-coloured boundary. Initial developments and experiments focus on pre-processing structured documents of which the layout, use of underlines, etc. is consistent through the documents.

The MONK system will subsequently be used to identify potential words, assigning a probability that a word is recognized in each region. The recognition accuracy can be improved by exploiting the recognized images. It is a tremendous advantage to be able to make use of multimodal combinations of text and images: the probabilities of structural elements in pictures are much more reliable (i.e., have a large raw count), than the probabilities for linguistic patterns that have a long-tailed distribution (occurring a few times). The underlying biological structure (has_wings, has_beaks), for example, creates common pictorial themes that can be exploited by infrastructure. These pictorial elements can be recognized separately and combined in a model using attribute learning which is an effective way to transfer knowledge between different labelled instances.

System II: Integration of Background Knowledge. This system functions as a bridge between bottom-up hypotheses that the computer generates from the illustrated handwritten pages and existing semantic bodies of knowledge. The bridge includes a mechanism which brings statistically inferred (deep learned) semantics in a fruitful dialogue with (formally coded) ontologies. While the labelling interface does not impose any limits on the labels, we can prompt users to favour terms from existing ontologies and vocabularies, using techniques such as auto-complete and by displaying extra context through synonyms and alternative spelling. The advantage of using standardized terms is that once words have been linked to standardized terms, background knowledge concerning the context of these terms can be used to bias the recognition of the scanned pages.

For instance, we know that most field note records from the NC collection are biodiversity observations. Consequently, we know that they contain a similar set of basic metadata required to describe such biodiversity observations. This includes (i) the name of the observed species, (ii) the researcher who identified it, (iii) the geographic location, (iv) the date of observation.

Many documents have a typical structure in which these metadata elements occur in specific positions. If MONK is uncertain about the term that corresponds with a word candidate, knowledge about the specific position and category that the multiple word candidates belong to, can help to make the right choice, especially for words that occur only a few times throughout the manuscripts. Consider the recognition of dates and place names. Drawings and sketches of zoological specimens often contain a toponym and the date of a sighting in the lower left corner (see Fig. 4). Place names are followed by a comma, and the month is represented in Latin characters followed by the year in Arabic

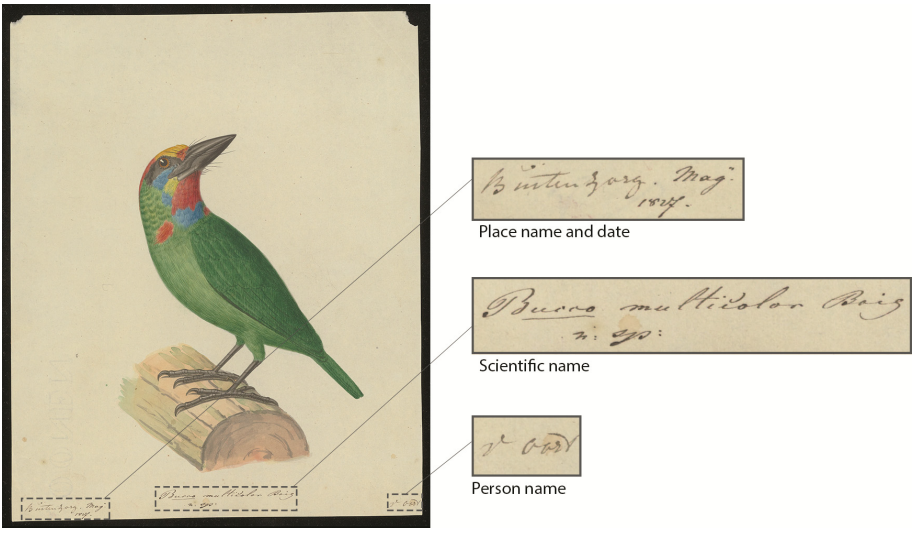


Fig. 4. Field drawing of a red-throated Barbet (*Begalaima mystacophanos)*, Buitenzorg, Java, May 1827. Figure created by Andreas Weber and Maarten Heerlien, licensed under CC-BY-SA 4.0. Drawing: Naturalis BC, NNM001000144_002.

digits. While the place and date are separate words, their proximity in a specific place on the page helps their identification as dates and place names [34]. Besides layout structure, ontological information such as hierarchical structures can be used to assist the word-recognition. Figure 4, for instance, was drawn in Java of which amongst others 'Buitenzorg' is a second level subdivision. Word recognition can use this information to bias the labelling towards 'Buitenzorg' instead of other, visually similar words.

The position and proximity of related words and ontological information can also be used to identify the names of observed species, which can be found in field drawings as well as in field notes. In field drawings, species names are noted at the bottom of the document (see Fig. 4). While species names typically consist of multiple words, these names follow the rules of (i.e., taxonomic) nomenclature. Regardless of the language of an individual field note, taxonomic names are represented according to the Linnean system of binominal nomenclature. It consists of a capitalized genus name, followed by a non-capitalized species name (e.g. *Burro multicolor Boie)*. Taxonomic names are written in italics in printed texts and are underlined in handwritten texts. Although the taxonomic classification of species may have changed since the early nineteenth century, these basic taxonomic categories can be used by MONK: order, family, genus, and species [35]. Species names will be the more useful, as they are always binomen, followed by author name, unless it is a new species (genus+species, (author)). If multiple adjacent word candidates possibly represent taxonomic terms, this knowledge can therefore be used to prioritize taxonomic names over other possible interpretations.

System III: Organization, Linking and Serving of Content. The accurate extraction of named entities such as person, species, and place names from handwritten illustrated collections is essential for making such collections available for researchers and a wider public. However, in order to enable algorithms to reason over the data to bootstrap or aid the word-recognition process, or to compare the extracted content with other resources, the data must be organized, structured and interlinked in a standard way. This will enable researchers to ask complex questions across the whole collection. Some projects, such as the Field Book Project,[3] already use standards for linking field books and other collections to their metadata, but none aim to link content. To do so effectively, we build on accepted community metadata standards, rather than developing a bespoke solution for this particular collection [36]. In this way, the work on the NC content serves as a proof of concept for other cultural heritage collections. The core resources we have adopted, for example, include the Darwin Core (DwC) standard for describing biodiversity data and the Dublin Core (DC) standard for describing data sources on the web.

The relationship between metadata entities is defined in an application ontology, represented in OWL,[4] which allows the semantic enrichment and integration of extracted archive content (e.g. species observations are made in a particular place, by a particular person). These relationships can be used to assist human and machine labelling as well as allowing the formulation of rich queries through the interface. For example, researchers could retrieve all bird species observed by a specific naturalist outside of

[3] http://biodivlib.wikispaces.com/The+Field+Book+Project, last accessed 2017/09/01.
[4] https://github.com/lisestork/NHC-Ontology, last accessed 2017/09/01.

Java, in South East Asia. The ontology also allows the consolidation of synonyms that are expressed with multiple different words or languages throughout the collection, such as place names, species names, or abbreviations, into single semantic concepts. Semantically integrated data will be stored in RDF (Resource Description Framework, the W3C standard for data interchange on the web), which will allow the content to be served as Linked Data, and more importantly, open the possibility of integration with other cultural heritage resources already available as Linked Data.

3 Discussion

This paper introduces a new digital infrastructure which allows curators of historical archives and manuscript collections to disclose and connect handwritten illustrated archives. Moreover, it presents initial results on the active learning performance of MONK on the NC collection, without the development of a prior language model. Owing to its visual approach, it is able to handle heterogeneous collections. Our use case offers a rich landscape of challenging visual material which we use to develop, synchronise and refine the infrastructure. In order to scale up MONK, the Making Sense project enriches it on three levels – recognition techniques, textual post-processing, and the querying of extracted results. On the level of recognition techniques we propose adapted methods for layout analysis, pattern recognition and machine learning. On the level of textual post-processing, we develop a system that functions as bridge between bottom-up hypotheses that the computer creates from the illustrated handwritten pages and existing semantic bodies of knowledge. This involves combining the use of ontologies and controlled vocabularies (for labelling text and images) and deep learning for image recognition. On the level of querying of extracted results, we semantically structure and link the extracted content in a standardized format, drawing on existing community standards and background knowledge to produce a resource that is interoperable with other collections. In sum: by combining handwriting recognition, layout segmentation, image recognition and ontological data annotation our infrastructure has the potential to significantly improve the automated extraction, classification and linking of knowledge from historical manuscript collections. It thus opens up new opportunities for scientific research, heritage institutions and publishers, while reducing the need for costly human intervention.

References

1. Heerlien, M., Van Leusen, J., Schnörr, S., De Jong-Kole, S., Raes, N., Van Hulsen, K.: The natural history production pine: an industrial approach to the digitization of scientific collections. J. Comput. Cult. Herit. **8**, 3:1–3:11 (2015)
2. Pethers, H., Huertas, B.: The Dollmann collection: a case study of linking library and historical specimen collections at the Natural History Museum, London. Linnean **31**, 18–22 (2015)
3. Ogilvie, B.: Correspondence networks. In: Lightman, B. (ed.) A Companion to the History of Science, pp. 358–371. Wiley (2016)
4. Ridge, M. (ed.): Crowdsourcing Our Cultural Heritage. Ashgate, Farnham (2014)

5. Franzoni, C., Sauermann, H.: Crowd science: the organization of scientific research in open collaborative projects. Res. Policy **43**, 1–20 (2014)
6. Terras, M.: Crowdsourcing in the digital humanities. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) A New Companion to Digital Humanities, pp. 420–438. Wiley, New York (2015)
7. Causer, T., Tonra, J., Wallace, V.: Transcription maximized; expense minimized? Crowdsourcing and editing The Collected Works of Jeremy Bentham. Lit. Linguist. Comput. **27**, 119–137 (2012)
8. Causer, T., Terras, M.: 'Many hands make light work. Many hands together make merry work': Transcribe Bentham and crowdsourcing manuscript collections. In: Crowdsourcing Our Cultural Heritage, pp. 57–88. Ashgate, Surrey (2014)
9. Orli, S., Bird, J.: Establishing workflows and opening access to data within natural history collections. Collections **12**, 147–162 (2016)
10. Mitchell, W.J.T.: Picture Theory: Essays on Verbal and Visual Representation. University of Chicago Press, Chicago (1994)
11. Kusukawa, S.: Picturing the Book of Nature: Image, Text, and Argument in Sixteenth-Century Human Anatomy and Medical Botany. University of Chicago Press, Chicago (2011)
12. Kwastek, K.: Vom Bild zum Bild - digital humanities jenseits des textes. In: Baum, C., Stäcker, T. (eds.) Grenzen und Möglichkeiten der Digital Humanities (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1) (2015)
13. van der Zant, T., Schomaker, L., Zinger, S., van Schie, H.: Where are the search engines for handwritten documents? Interdisc. Sci. Rev. **34**, 224–235 (2009)
14. Mühlberger, G.: Die automatisierte Volltexterkennung historischer Handschriften. In: Digitalisierung im Archiv: Neue Wege der Bereitstellung des Archivguts, pp. 87–116. Archivschule Marburg, Marburg (2015)
15. Schomaker, L.: Design considerations for a large-scale image-based text search engine in historical manuscript collections. Inf. Technol. **58**, 80–88 (2016)
16. Mees, G., van Achterberg, C.: Vogelkundig onderzoek op Nieuw Guinea in 1828. Zoologische Bijdragen **40**, 3–64 (1994)
17. Klaver, C.J.: Inseparable Friends in Life and Death: The Life and Work of Heinrich Kuhl (1797–1821) and Johan Conrad van Hasselt (1797–1823). Barkhuis, Groningen (2007)
18. Temminck, C.J., Müller, S., Schlegel, H., de Haan, W., Korthals, P.W.: Verhandelingen over de natuurlijke geschiedenis der Nederlandsche overzeesche bezittingen. Luchtmans, Leiden (1839–1847)
19. Roberts, T.R.: The freshwater fishes of Java, as observed by Kuhl and van Hasselt in 1820-23. Zoologische Verhandelingen **285**, 1–93 (1993)
20. Fransen, C.H.J.M., Holthuis, L.B., Adama, J.P.H.M.: Type-catalogue of the Decapod Crustacea in the collections of the Nationaal Natuurhistorisch Museum, with appendices of pre-1900 collectors and material. Zoologische Verhandelingen **311**, 1–344 (1997)
21. Hildenhagen, T.: Heinrich Kuhl - Das Leben eines fast vergessenen Naturforschers aus Hanau. Neues Magazin für Hanauische Geschichte, pp. 110–214 (2013)
22. See for instance the digital Cyclopaedia of Malaysian Collectors. http://www.nationaalherbarium.nl/FMCollectors/Introduction.htm. Last Accessed 08 Sep 2017
23. Hoogmoed, M.S., Gassó Miracle, M.E.: Type specimens of recent and fossil Testudines and Crocodylia in the collections of NCB Naturalis, Leiden, the Netherlands. Zoologische Mededeelingen **84**, 159–199 (2010)
24. van der Zant, T., Schomaker, L., Haak, K.: Handwritten-word spotting using biologically inspired features. IEEE Trans. Pattern Anal. Mach. Intell. **30**, 1945–1957 (2008)

25. Van Oosten, J.-P., Schomaker, L.: A Reevaluation and benchmark of hidden Markov models. In: 2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 531–536 (2014)
26. Van Oosten, J.-P., Schomaker, L.: Separability versus prototypicality in handwritten word-image retrieval. Pattern Recognit. **47**, 1031–1038 (2014)
27. READ project website. https://read.transkribus.eu/. Last Accessed 27 July 2017
28. He, S., Wiering, M., Schomaker, L.: Junction detection in handwritten documents and its application to writer identification. Pattern Recognit. **48**, 4036–4048 (2015)
29. Günter, S., Bunke, H.: HMM-based handwritten word recognition: on the optimization of the number of states, training iterations and Gaussian components. Pattern Recognit. **37**, 2069–2079 (2004)
30. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
31. Graves, A.: RNNLIB: a recurrent neural network library for sequence learning problems. http://sourceforge.net/projects/rnnl/. Last Accessed 01 Sep 2017
32. Bulacu, M., Brink, A., van der Zant, T., Schomaker, L.: Recognition of handwritten numerical fields in a large single-writer historical collection. In: 2009 10th International Conference on Document Analysis and Recognition, pp. 808–812 (2009)
33. Yan, K., Verbeek, F.J.: Segmentation for high-throughput image analysis: watershed masked clustering. In: Margaria, T., Steffen, B. (eds.) ISoLA 2012. LNCS, vol. 7610, pp. 25–41. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34032-1_4
34. Shi, Z.: Handwritten document images based on positional expectancy, Master thesis, Artificial Intelligence, University of Groningen, the Netherlands, May 2016
35. Gassó Miracle, M.E.: On whose authority? Temminck's debates on zoological classification and nomenclature: 1820–1850. J. Hist. Biol. **44**, 445–481 (2011)
36. Stork, L., Weber, A.: A linked data approach to disclose handwritten biodiversity heritage collections. In: Presented at the Digital Humanities Benelux Conference 2017 (2017)