From Historical Handwritten Manuscripts to Linked Data

Lise Stork[1,2]✉, Andreas Weber[3], Jaap van den Herik[1,2],, Aske Plaat[1,2], and Fons Verbeek[1,2] 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 for Data Science, Leiden, the Netherlands
h.j.vandenherik@law.leidenuniv.nl
[3] University of Twente, Twente, the Netherlands,
a.weber@utwente.nl

Abstract. Museums, Archives, Libraries and other institutes, specifically those in the cultural heritage domain, make increasing use of Semantic Web technologies to enrich and publish their collection items. Fewer cases exist where the *contents* of those items are also enriched using similar methods, disclosing the details contained within historical handwritten manuscripts. We argue that the enrichment of historical manuscripts is of central importance to the disclosure of cultural heritage archives. Elucidating the contents of historical manuscripts is, however, a time-consuming process that requires domain expertise. Different workflows have therefore been proposed to accelerate and improve this process. In this study, we present an analysis of different approaches, focussing specifically on the provenance requirements for annotating and interpreting historical manuscripts so that the contents can be published online as FAIR (Findable, Accessible, Interoperable and Reusable) data. Furthermore, we argue that provenance can play a central role in quality assessment. We demonstrate our findings with a case study from the natural history domain, where we have developed a semantic framework for extracting, annotating and curating regions of interest from digitised handwritten, historical manuscripts.

Keywords: Linked Data · Cultural Heritage · Provenance · Handwritten Manuscripts · Crowdsourcing · Semantic Annotation · Regions of Interest · Natural History

1 Introduction

Many digital libraries contain handwritten manuscripts from various domains. They are physically stored in museums and other institutions all over the world. The main challenge is the transformation of a collection of digitised manuscripts to a searchable knowledge base. This is especially difficult when dealing with historical texts containing multiple languages, hard-to-read handwriting and historical content. Challenges such as name ambiguity and data interpretability seem

to ask for more 'intelligent' web technologies. The promising features that semantic web technologies offer are expected to cover these challenges well. Serving data to the web of Linked Data has many benefits: (i) the adoption of Internationalised Resource Identifiers (IRIs) for the representation of named entities enable more accurate content descriptions, (ii) data become interoperable, (iii) through Semantic Web services, data can be federated and integrated between distributed collections [12].

Digital libraries, especially from the cultural heritage domain, have recently become a major application for Linked Data technologies. Various cultural heritage initiatives already enrich their data on a collection level and on an item level using the Linked Open Data principles[1] [5, 7, 21, 22]. Some institutes even model interesting *named entities* such as locations and persons, mentioned in their collections, using dereferenciable URIs from, for instance, *Geonames* and International Authority Files such as *VIAF*.[2] A telling example is the Dutch Ships and Sailors project [3], which resulted in four curated datasets on Dutch maritime history in the form of Linked Open Data. Another example is the Europeana project,[3] an initiative that operates on an international scale, offering rich semantic representations of cultural heritage items that are utilisable by other institutions.

Only a few technological infrastructures, however, describe the Regions Of Interest (ROIs) in images from cultural heritage archival collections in a similar way. Annotation of the textual content helps museums and other cultural heritage institutions to disclose their manuscript collections. Annotation is done using crowdsourcing [18] or nichesourcing [4] - harvesting labels from handwritten manuscript images using the crowd or domain experts -, automated processes such as handwriting recognition or word spotting, or a combination thereof. Often, annotations are formatted in XML using a specified format, such as the Text Encoding Initiative (TEI) scheme.[4] While models and tools to annotate web resources using the principles of Linked Data exist [17], they are, to the best of our knowledge, seldom applied to annotating ROIs in historical handwritten manuscripts. As a result, a small number of collections make use of the full potential of Linked Data: data are stored in a variety of formats and are usually not publicly available for reuse by other researchers. Moreover, even fewer cases exist where ROIs are actually semantically enriched with IRIs from the Linked Open Data cloud and interlinked within and beyond the current collection.

Using Linked Data to annotate manuscripts helps publish the content as FAIR (Findable, Accessible, Interoperable, Reusable) data [26], which requires annotations to be stored in a transparent way. The provenance of annotations needs to be tracked on multiple levels, for instance, regarding the annotation process and regarding the location of the annotated text in the digital images.

[1] https://www.w3.org/wiki/LinkedData
[2] https://viaf.org
[3] https://www.europeana.eu/portal/
[4] http://www.tei-c.org/index.xml

Annotation provenance can be used for annotation quality control, attribution of annotation efforts and the stimulation of scientific discourse. Current information systems usually only capture minimal provenance information [9]. A multitude of requirement descriptions and W3C[5] recommended vocabularies exist; however, they still need to be implemented in many archives and collections. Methodologies and principles to incorporate annotation provenance in the design and implementation of software systems are lacking: clear requirements and in-use examples are required [9].

In this study, we analyse different workflows for the enrichment of historical handwritten manuscripts, in section **2**, and recommend a workflow for expert data curation of a specific type of manuscript. We subsequently assess which provenance of annotation objects should be tracked during the enrichment process, which is described in section **3**. We then discuss how this provenance can potentially be used in practice, for instance to measure and subsequently safeguard data quality. Lastly, in section **4**, we demonstrate an implementation of this approach using a semantic annotation tool for natural history field books, the *Semantic Field Book (SFB) Annotator*[6] that, alongside semantic information, stores annotation provenance.

2 Manuscript enrichment workflows

To enrich the content of manuscripts, different workflow methods have been developed for transforming digital images or texts to a database. Most, however, generate flat or semi-structured text files, rather than Linked Data. Here we discuss a set of workflows that enrich digitised manuscripts. The meaning of terms may vary between studies, so we distinguish (1) *verbatim transcriptions*: translations from texts in digital images to character encodings, (2) *labels*: mappings from a ROI in an image to computer text, usually a word-zone to a textual label as input to a learning system, and (3) *annotations*: notes or comments added to an image or text; they point to a specific region in a document and provide metadata about the verbatim transcription such as its semantic type or, for instance, the annotation process. Figure 1 graphically presents a variety of workflows for the enrichment of documents. The workflows vary in (i) the richness with which they describe the content, (ii) the percentage of text that is actually transcribed and (iii) the format in which the output is presented. In the coming sections we describe some initiatives that follow these different workflow methods.

Workflow 1. Word-zone labelling. 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]. With their system they have, amongst others, enriched a set of letters of *Joaquim Nabuco*, one of the key figures in the campaign for freeing black slaves in Brazil (1861-1910).

[5] https://www.w3.org/Consortium/mission
[6] https://github.com/lisestork/SFB-Annotator/

Another example is *Transkribus*, a platform for the enrichment of historical documents. A user can label sentences which are then used for training using HTR [14]. This project implements a form of semantic enrichment: labellers can flag certain named entities, such as, locations or persons with a predefined or user created tag set. These tags are, however, not defined in a sharable format.

Texts need not necessarily be fully transcribed. The aim of the *MONK* handwriting recognition system [20] is not full-transcription per se, but rather searchability of the content. The system is adaptive and therefore the labelling, either by a human or an machine, is more targeted: once certain words are labelled, the machine starts searching for these words in other parts of the manuscript collection. The system has already processed many documents, such as the *Dead Sea Scrolls*; Hebrew manuscripts encountered in the Qumran Caves near the Dead Sea.

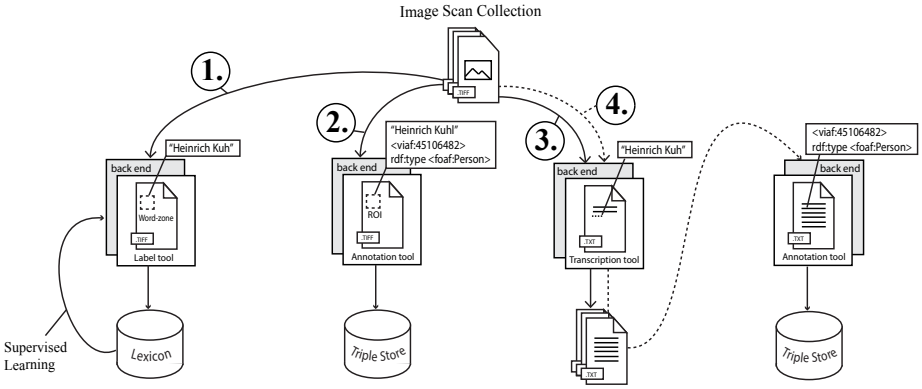

Fig. 1. Manuscript enrichment workflows

Workflow 2. Semantic annotation of digital images. Accurator[7] 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. 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 [6].

The *sharedCanvas* collaborative model also uses building blocks of Linked Data: Open Archive Initiative Object Exchange and Reuse (OAI-ORE) aggregations and Open Annotation Collaboration (OAC) annotations [11], to enrich image and text layouts *within* medieval manuscripts. The initiative uses these to enrich five different cultural heritage data sets [19].

Finally, the *SFB-Annotator*[6] [23], semantically annotates the content of digitised manuscripts from natural history collections using an application ontology and the Web Annotation Data Model.

[7] http://www.accurator.nl/

Workflow 3. Full-text transcription. Museums, archives and libraries around the world are beginning to notice the potential of crowdsourcing [18]. Hence, examples of initiatives that fully transcribe manuscripts manually are plentiful. The *Field Book Project*, a project set up by the Smithsonian Institution Archives in collaboration with the National Museum of Natural History. The project uses the crowd to harvest full-text transcriptions from field books [1]. Another example is the *Transcribe Betham* initiative that will digitise and, also via crowdsourcing, transcribe 12,500 folios from jurist Jeremy Bentham (1748-1832), stored in the University College London digital archive, through a media-wiki interface [16].

Workflow 4. Semantic annotation of text. Kiryakov et al. [15] discuss the principle of semantic annotation of texts. Their infrastructure is based on the annotation of text files that exist in databases on the web. Its technology can be used as part of a workflow such as is visible in figure 1, workflow **(4)**. The authors very clearly define semantic annotations as assigning semantic descriptions to entities in a text. Documents are then indexed and thus retrieved with respect to their semantic type instead of their keywords.

Annotea[8] is a shared web annotation system which is, just as the aforementioned technology, based on the semantic annotation of text files that originated on the web. In the Annotea architecture, annotations exist externally from the documents on *annotation servers*. The system uses XPointer[9] to let an annotation point to a piece of text. Other users are able to add their annotations to the annotation servers. Annotea makes use of RDF and HTTP with the aim to use as much existing W3C specifications as possible [13].

The project *From Documents To Datasets* [24] is an example of the full workflow **(4)** from figure 1. They first fully transcribe field book manuscripts after which they semantically enrich the transcribed text. Instead of an application ontology, they use a template.

Full-text transcription provides a solution for institutes to disclose their digital manuscript collections, but it is a time-intensive procedure that outputs flat or semi-structured text files rather than structured data. This means that one processing step is still required to transform the transcripts into enriched data. We opt for a more targeted approach where interesting regions of content are directly labelled and semantically enriched through a nichesourcing initiative. Experts in this case decide which regions are interesting to answer their research questions. Although some extra work is required to semantically annotate texts with Linked Data, the omission of the full-transcription step saves time as a smaller amount of text needs to be transcribed; preferably only *named entities* are transcribed and semantically annotated. This does not mean that other ROIs cannot be labelled and annotated, but this is (in most cases) not required to construct rich semantic queries or aggregate informative content across cultural heritage collections. Moreover, knowledge bases can be pre-populated with

[8] https://www.w3.org/2001/Annotea/
[9] https://www.w3.org/TR/2003/REC-xptr-framework-20030325/

background knowledge, such as relevant locations from the Geonames database or relevant persons from the VIAF authority IRIs, helping annotators to use the correct named entities for annotation. Using Linked Data for annotation helps remove ambiguity as IRIs contain rich descriptions. The name *'Heinrich Kuhl'*, for instance, is ambiguous as multiple individuals carry that name. If we instead label with the IRI `https://viaf.org/viaf/45106482/`, we can agree that the content concerns Kuhl, Heinrich (1797-1821), a german zoologist.

3 Provenance

Provenance refers to the origins of data: how were they produced, by whom, in what way and why. Initiatives are set up globally that describe provenance frameworks, vocabularies and requirements for publishing data on the web. Groth et al. [9] describe provenance dimensions for storing provenance of web content and how this provenance can be used in practice. The dimensions were constructed from user requirements within different domains. The Web Annotation Data Model[10] adheres to many of the requirements and is a W3C recommendation. It models provenance of multimedia annotations on the web and its vocabulary is written in RDF. It is therefore suitable for the description of annotations that link to digitised manuscript images [10, 11]. From the provenance dimensions as described by Groth et al., we discuss those required for annotating cultural heritage manuscripts in section 3.1, and describe how these can be used in practice in section 3.2.

3.1 Content

The emphasis of this study is on provenance relating to the *annotation object*, and on how its provenance can be used for quality control and the stimulation of academic discussions about the content. We therefore add one dimension here for the description of the content: the ROI coordinates and the source image to which an annotation is added, based on the Web Annotation Data Model.

Object The object in this case is the annotation, while in this case conceptual, it serves as glue to connect all annotation related provenance.

Target Target provenance links the annotated digital image, or ROI within a digital image, to the annotation. Extracted information can hereby be traced back to its primary source and location.

Attribution Attribution provenance refers to sources or entities that contributed to the creation of the annotation, such as the annotation tool and *creator* of the annotation. In crowd- or nichesourcing efforts, the creator usually is a human. In cases where annotation is performed by automated recognisers, the creator can also be a software system.

Process Provenance relating to the annotation process is retained. Its motivation, the timestamp of its creation and the tool with which it was created and its corresponding version.

[10] https://www.w3.org/TR/annotation-model/

Justification Recording how and why a particular decision is made is important for, for instance, a correct interpretation of the transcription or semantic type of the annotation.

3.2 Use

Combined with some of the provenance dimensions listed by Groth et al., we discuss one extra dimension: how provenance of annotations permits a user to use annotation provenance for research or data quality control.

Data quality Where a domain ontology and semantic model is used for annotation, researchers know what constitutes a full annotation record and can therefore assess the completeness of the annotated data using ROI provenance. When a base element of a record is annotated, e.g. the unique identifier, all instances of ontology classes in a complete record can be instantiated and linked together. One scenario where this could occur is the annotation of medieval manuscripts describing legal acts: an application ontology would formally define the basic logical components in the legal acts and their links, described by van Eck and Schomaker [8]. Instances and links of the application ontology can be instantiated with the annotation of the word *Item*, indicating the start of a new act. During a semantic annotation effort, annotations are linked to the instantiated entities. Non-annotated entities are elucidated as they have not been linked, via annotations, to the digital image. Section 4 demonstrates this principle.

Trust judgements Using attribution of annotations, judgements can be made regarding the quality of annotations. Annotation profiles can be maintained, using for instance authority files such as the Virtual International Authority File (VIAF[11]) or ORCID identifiers.[12] Records can be maintained specifying the amount of annotations created per annotator, the percentage of these being validated and the domains in which they operate.

Scholarly discussions Historical manuscripts need to be interpreted. In many cases, there are multiple possible interpretations. A provenance model that allows assertions about the content to exist in parallel will therefore stimulate scholarly discussions. For example, in cases where the original text is ambiguous, illegible, or contains difficult historical contexts, different semantic interpretations may exist in parallel.

4 Case study: from *field books* to Linked Data

Below we describe a case study taken from the Making Sense project.[13] Within that project, methods are being developed for automated semantic annotation of natural history collections [25], in order to accelerate the process of making

[11] https://viaf.org
[12] https://orcid.org
[13] makingsenseproject.org

such resources accessible to the natural and cultural history research community. Our use case consists of 8000 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. An ontology[14] and web application[6] have been realised to enable fine-grained annotation of ROIs in field book manuscripts. Using the Web Annotation Data Model, a ROI is linked, via an annotation object, to a specified semantic type from the ontology. Provenance regarding the attribution, process and justification is attached to the annotation object. A sample set of annotations from one field book is available online via a SPARQL endpoint.[15]

4.1 Workflow

The project adopts workflow 2 from figure 1, the semantic annotation of ROIs in digital images. We argue that this workflow is especially well suited for manuscripts where the content is structured and contains recurring elements. Species observation records from field books are highly structured; they each contain a set of logical components: a taxonomical name, a geographical location and a person, but also anatomical entities and species characteristics. We call the set of logical components, or *named entities*, connected by properties, a record. Semantically annotating such records is more straightforward than in the case of free text. The named entities are recurring and usually salient, meaning that they take up a prominent position on the page and are highlighted by means of textual formatting: underlined, bold, italic, enumerated and other.

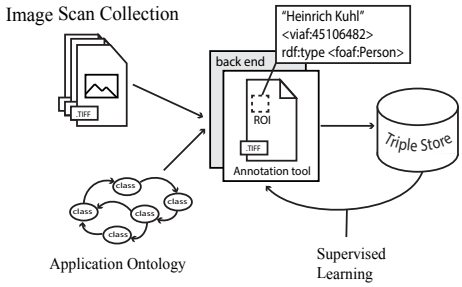

Fig. 2. A workflow for semi-automated semantic annotation of salient entities in digitised manuscripts

Full-text transcription is time-intensive and produces a lexicon of which a big percentage will most likely not be used as keywords to search the archive. Using workflow 2, a quick representation of the archive is built that allows rich semantic queries. For an in-depth analysis of the documents, the users can quickly be pointed to interesting pages, for instance all pages that contain a taxonomical

[14] https://github.com/lisestork/NHC-Ontology
[15] (http://makingsense.liacs.nl/rdf4j-server/repositories/NC

name that has as a scientific author *Étienne Geoffroy Saint-Hilaire*.

Our final goal is to combine workflows 1 and 2, see figure 2. Expert curated semantic annotations are now being used as input to a train supervised learning system, in order to recognise the named entities automatically. In combination with an HTR and Layout Analysis system such as MONK, semi-automated semantic annotation can be realised.

4.2 Linked Data annotations

Semantic annotation using Linked Data is recommended for the correct interpretation of the complex content of historical field book manuscripts. *Buitenzorg*, referred to in the manuscripts, is a historical name which is, using Linked Data from the Geonames database, easily georeferenced to its current name *Bogor*. Furthermore, biological taxonomical names are ambiguous due to evolving taxonomical systems and need to be described using Linked Data and referenced carefully to current taxa. Cases exist where name duplicates refer to different organisms - for instance, where one refers to a mammal and the other to a plant. Discussion concerning their interpretation in this case is highly appreciated. The scientific publisher of a taxonomical name is often decisive for its interpretation. In our use case, the word *'nobis'* is used by naturalists, meaning *'by us'*. Semantically, this name can be annotated as a person name whereas linguistically, it is an indirect object. This type of information is not captured using full-text transcription.

4.3 Data quality

In the methodology that we adopt, all 'base' elements of a record are instantiated after the annotation of the 'start' of the record, as discussed in section 3.1. In this case, the taxonomical name indicates the start of a record and the organism receives a unique identifier. When an element from a record is not annotated, no link exists between the instance and the ROI in the manuscript. Visualisations show these missing links and can help discover missing information. The graph in figure 3 shows a visualisation of an annotated record. However, only two elements are annotated: the taxonomical name *'Pteropus minimus'* and the scientific author *'Étienne Geoffroy Saint-Hilaire'*. This is far below the minimum information we would expect for a species observation. A date and a location, for example, are also expected. Some instances from the ontology, such as the *nc:organism1*, are not explicitly found in the manuscripts and serve as glue to logically connect all instances. By querying for missing information, curators of this collection can determine if a record is incomplete due to poor annotation, or due to missing information in the original source. A query can be constructed that retrieves all instantiated named entities, for example all locations, without a link to a manuscript. The result will point a user to the records that miss annotations. One can go back to these observation records and verify the content. Some annotations may have been omitted due to ambiguity. For example, if an annotator cannot interpret the historical name for a location.

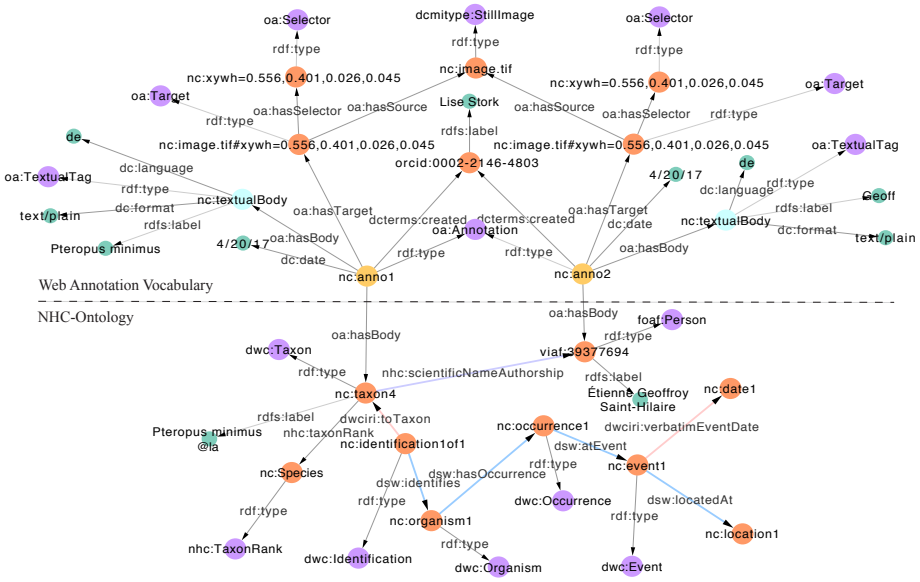

Fig. 3. Part of an annotated observation record

5 Conclusions and future work

In this study, we have analysed a set of workflows used to enrich digitised manuscripts. We have argued that, although full-text transcription is an effective procedure that is often used, it is a very time-consuming process that does not scale well. Transcriptions are, furthermore, presented as unstructured or semi-structured text which still require enrichment. Tools should be developed that facilitate semantic annotation of digitised manuscripts. Storing provenance of these annotations is key to publishing the data online as FAIR data, making them reusable by other cultural heritage institutions. Many provenance recommendations have been proposed for web-annotated content. In this study, we have described the information that is required for publishing Linked Data annotations of historical manuscript content, and opt for the use of one W3C recommendation, the Web Annotation Data Model. This model is being adopted to annotate digital images on the web, contributing to the web of Linked Data. We have furthermore described a specific workflow in the context of our use case. We demonstrated how storage of annotation provenance can help give insight into the quality of the data.

As we move forward to combine workflows 1 and 2 in the Making Sense project, we anticipate a need to extend and enrich our provenance model. We will incorporate HTR or other automated processes and will therefore need to capture the provenance of those automated processes, such as, changing data, process parameters, algorithm versions and data transformations. In future work we will present an in-depth analysis of how annotation provenance on the web can

stimulate scholarly discussions concerning the correct interpretation of named entities in historical handwritten texts by allowing annotations to exist in parallel.

To conclude, we demonstrate that by directly semantically annotating named entities from the content of digitised historical manuscripts, and publishing them online as Linked Open Data, the quality of the annotations can be assessed, and 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 (2010), https://siarchives.si.edu/about/field-book-project, https://siarchives.si.edu/about/field-book-project, accessed: 14-04-2018
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. de Boer, V., van Rossum, M., Leinenga, J., Hoekstra, R.: Dutch ships and sailors linked data. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) International Semantic Web Conference (ISWC 2014). Lecture Notes in Computer Science, vol. 8796, pp. 229–244. Springer International Publishing, Cham (October 2014)
4. De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., Schreiber, G.: Nichesourcing: harnessing the power of crowds of experts. In: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. (eds.) International Conference on Knowledge Engineering and Knowledge Management. Lecture Notes in Computer Science, vol. 7603, pp. 16–20. Springer, Heidelberg, Berlin (2012)
5. De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., Schreiber, G.: Supporting linked data production for cultural heritage institutes: The amsterdam museum case study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) The Semantic Web: Research and Applications. ESWC 2012. Lecture Notes in Computer Science, vol. 7295, pp. 733–747. Springer, Berlin, Heidelberg (2012)
6. Dijkshoorn, C., De Boer, V., Aroyo, L., Schreiber, G.: Accurator: Nichesourcing for cultural heritage. arXiv preprint arXiv:1709.09249 (2017)
7. Dragoni, M., Tonelli, S., Moretti, G.: A knowledge management architecture for digital cultural heritage. Journal on Computing and Cultural Heritage (JOCCH) **10**(3), 1–18 (August 2017)
8. Ritsema van Eck, M., Schomaker, L.: Formal semantic modeling for human and machine-based decoding of medieval manuscripts. In: Proceedings of Digital Humanities. pp. 336–338. University of Hamburg (July 2012)
9. Groth, P., Gil, Y., Cheney, J., Miles, S.: Requirements for provenance on the web. International Journal of Digital Curation **7**(1), 39–56 (2012)
10. Haslhofer, B., Sanderson, R., Simon, R., Van de Sompel, H.: Open annotations on multimedia web resources. Multimedia Tools and Applications **70**(2), 847–867 (2014)
11. Haslhofer, B., Simon, R., Sanderson, R., Van de Sompel, H.: The open annotation collaboration (oac) model. In: Workshop on Multimedia on the Web (MMWeb 2011). pp. 5–9. IEEE Computer Society, Washington, DC, USA (September 2011)

12. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web. Synthesis Lectures on the Semantic Web: Theory and Technology **2**(1), 1–159 (2012)
13. Kahan, J., Koivunen, M.R., Prud'Hommeaux, E., Swick, R.R.: Annotea: an open rdf infrastructure for shared web annotations. Computer Networks **39**(5), 589–608 (2002)
14. 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)
15. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D.: Semantic annotation, indexing, and retrieval. Journal of Web Semantics **2**(1), 49–79 (December 2004)
16. Moyle, M., Tonra, J., Wallace, V.: Manuscript transcription by crowdsourcing: Transcribe bentham. Liber Quarterly **20**(3-4) (2011)
17. Nixon, L., Troncy, R.: Survey of semantic media annotation tools for the web: towards new media applications with linked media. In: European Semantic Web Conference. pp. 100–114. Springer (2014)
18. Oomen, J., Aroyo, L.: Crowdsourcing in the cultural heritage domain: opportunities and challenges. In: Proceedings of the 5th International Conference on Communities and Technologies. pp. 138–149. ACM (2011)
19. Sanderson, R., Albritton, B., Schwemmer, R., Van de Sompel, H.: Sharedcanvas: a collaborative model for medieval manuscript layout dissemination. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. pp. 175–184. ACM (2011)
20. Schomaker, L.: Design considerations for a large-scale image-based text search engine in historical manuscript collections. It - Information Technology **58**(2), 80–88 (April 2016)
21. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Osenbruggen, J., Tordai, A., et al.: Semantic annotation and search of cultural-heritage collections: The multimedian e-culture demonstrator. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4), 243–249 (2008)
22. Steiner, E., Koch, C.: A digital archive of cultural heritage objects: Standardized metadata and annotation categories. New Review of Information Networking **20**(1-2), 255–260 (2015)
23. Stork, L., Weber, A., Miracle, E.G., Verbeek, F., Plaat, A., van den Herik, J., Wolstencroft, K.: Semantic annotation of natural history collections, under review
24. 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 (July 2012)
25. Weber, A., Ameryan, M., Wolstencorft, K., Stork, L., Heerlien, M., Schomaker, L.: Towards a digital infrastructure for illustrated handwritten archives. In: Loannides, M. (ed.) Digital Cultural Heritage. Information Systems and Applications, incl. Internet/Web, and HCI, vol. 10605, pp. 155–166. Springer International Publishing (April 2018)
26. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)