https://doi.org/10.2352/issn.2168-3204.2019.1.0.20

Automatic metadata (entity) extraction and workflow efficiency: real life solutions

Martijn van der Kaaij, Heron Information Management LLP, Weert, The Netherlands

Abstract

In a changing digitization landscape, automatic metadata extraction is becoming more important than ever before. At the same time, requirements regarding the extracted metadata are becoming more demanding as well: we are no longer interested in just extracting some data, we want to extract and identify entities.

Large leaps have been made on different aspects of metadata extraction, however integrated and effective workflows successfully and efficiently applying metadata extraction to real collections in a market environment are still rare.

This paper describes the research, the principles applied and the implementation of just such a workflow.

The need for automated metadata extraction in a changing field

At this moment in time, there is not much profit in digitization of cultural heritage material, as prices paid per object or page are low, and a lot of effort must usually be spent on adding high quality descriptive metadata. Therefore, any attempt to increase the amount of metadata that can be extracted automatically is worthwhile, especially if the extracted data is validated properly (and immediately).

Furthermore, the nature of digitization projects is changing, from scanning large amounts of more or less uniform data to digitizing smaller, more diverse collections, where 'digitizing' means not just scanning, but also enriching the data and fitting it in existing metadata structures. Increasingly, the creation of metadata objects suitable for a linked data universe is a part of project specifications.

To deal with this change in a cost effective way, automatic metadata extraction is becoming even more vital.

The problem: under use of metadata in digitization workflows

Figure 1 shows a mass digitization workflow as it usually presents itself in the projects undertaken by our company. The drawing is based on the IDEF0 method [1]. Circles for start and end points, white boxes for process steps, with yellow boxes on top for constraints (or 'control', e.g. time, available resources, standards) and black and black boxes below for resources (or 'mechanism', e.g. divisions, people, roles, systems). The 'data provider' attached to the first and last step can be the actual owner of the resources to be digitized, but it can also be a party overseeing the digitization process for one or more owners. In many of 'our' projects the National Library of the Netherlands would take this role. The 'digitization service provider' is the main contractor: parts of the work may be subcontracted to other companies.

Offering services regarding quality control to both data providers and digitization service providers, we soon became aware of the bottlenecks in the workflow.

Firstly, the quality and extent of the inventory prepared in the first step is very dependent on the local metadata resources of the data provider. For example: if an archival data provider was used to working with the Encoded Archival Description (EAD), we could expect high quality inventories, but on other occasions we just had to be happy with a list of inventory numbers. When dealing with library data, a lot would depend on the quality of the cataloging records on which the inventory would be based. This resulted in a situation where the inventory was mainly good for counting in the last step of the workflow: did we end up with the expected amount of digital

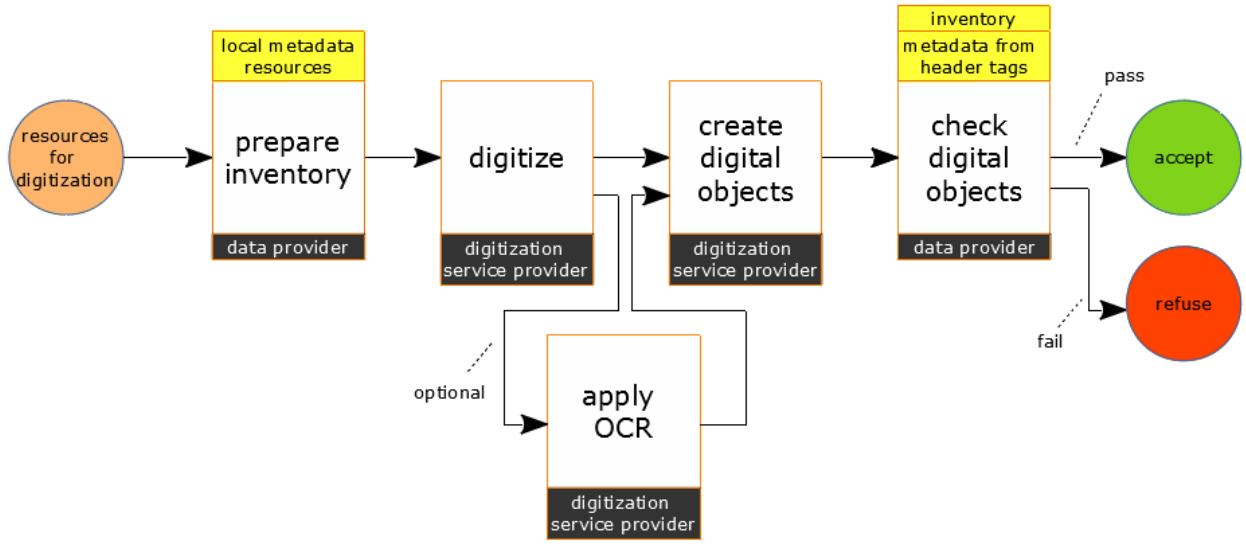

Figure 1. *Initial mass digitization process*

objects at the end? Even if good descriptive metadata were present, they were not used in the workflow.

Secondly, OCR (Optical Character Recognition) would only be applied when processing printed material, as Handwritten Text Recognition (further: HTR) is still in its infancy. When OCR was applied, the resulting data would hardly play any part in the digitization process. It would just be stored as part of the digital subjects and used for feeding access services later on.

Thirdly, in some projects metadata would be extracted from header tags, but this would only serve limited functions, for example to relate scans to target images on model, make and timestamp.

Having designed document workflows in commercial environments for years, and having seen the use of metadata extracted from the documents in these workflows, the mass digitization workflow definitely had an unfinished feel to it. However, we couldn't do much with that feeling: as a commercial company, in most of its projects Heron Information Management cannot usually afford time to research these issues or run experiments. In early 2018, however, an opportunity arose to do just that. LOTS imaging [2], a new digitization company on the Dutch market was interested in providing tailor made solutions, using our work on workflow optimization and the accompanying software. As they wanted to be sure to offer innovative services that were beneficial to their customers, they thought it necessary to spend time on exploring new approaches. Their flexible business model allowed them to team up with us to find the space for these new approaches.

As subject matter for our efforts, we chose a corpus of notary acts dating from around 1800, as these documents are fairly structured when it comes to the place and the way key items like dates and names are being recorded.

Solutions

Having set up our partnership, we needed to define the specific goals of our project (in order of priority, highest first):

- automatically extract key descriptive elements
- standardize the extracted metadata (if applicable)
- 'resolve' the extracted metadata to existing entities, or create new entities

As for the descriptive elements, we decided to target dates, place names, persons and subjects first. Looking at the selected corpus, the locations where this information can be found in the documents can be predicted quite accurately.

An example of standardization would be the translation of dates from the French Revolutionary Calendar, which is used in part of the corpus, to the Gregorian Calendar.

The wish to resolve the extracted metadata to entities is a result of the desire mentioned above to do more than return digital images. Where, in the traditional workflow, the descriptive metadata to access the digitized documents had to be generated form the OCR-ed text at a later stage, our new style projects will deliver ready to use linked data objects as output of the digitization phase.

Work on the project took place in different areas:

HTR

First, we addressed the matter of HTR, or, to be more specific: the recognition of certain repeated and/or related words or phrases in a piece of handwritten text. Two fundamentally different approaches were considered. First, we looked into "word graph based keyword spotting" [3], which, while being applied mostly to spoken text, has also been used to process some interesting written text collections [4][5]. Although very interesting and, in some respects, the more scientific way forward, we did not see how we could apply this method quickly to an actual workflow within the context of our project.

Looking further, we encountered the Transkribus project [6]. Central to Transkribus (from our point of view) was the possibility to set up and 'train' an HTR model. For a corpus like ours, with a limited set of hands, a fairly uniform document structure and a limited, quite clearly defined vocabulary, such an approach seemed very promising. Although Transkribus itself is not meant for a commercial environment like ours, we did apply a lot of its principles to our project.

Image processing tools

Having already developed methods in the past for setting 'masks' on digital images to define the most likely areas to find, for example, bar codes, and having developed tools for cropping, splitting and stitching digital images which also involve quite a lot of analysis of the content of the image, we could fairly easily set up similar tools for our new workflow. This means that, if full OCR is not required, it is possible to automatically identify and select only those areas of the document that need OCR-ing to acquire the required data.

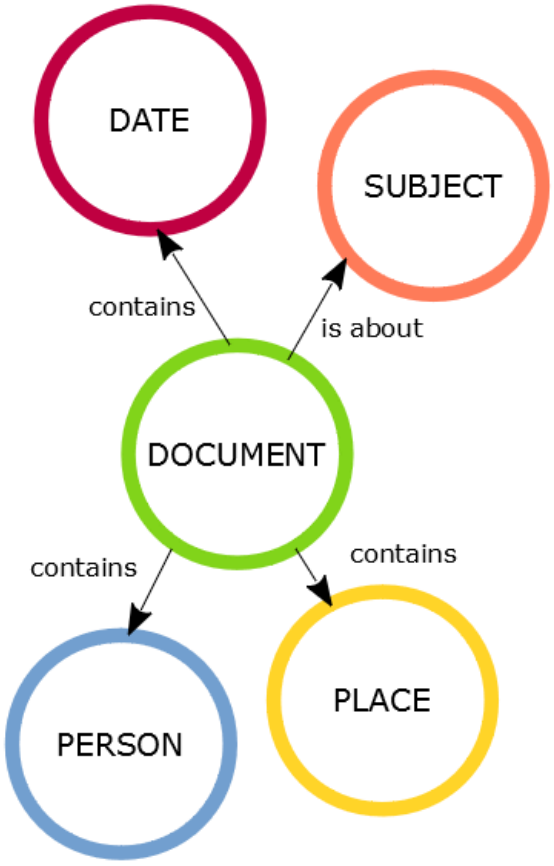

Figure 2. *Initial ontology for notary acts*

Entity resolution an ontology set up

Strictly speaking, identifying a date is already resolving an entity, as, in a linked data context, dates are entities. However, true added value would come form harvesting entities representing persons and subjects from the corpus. Therefore, we needed to create an ontology for our corpus.

As can be seen in Figure 2, the initial ontology was quite basic. It recognizes five entities (the circles). Each entity has one or more attributes ("first name" for a person, for example), but they are not shown in the illustration.

As for subjects, we would usually be able to establish them, but with people, places and dates, things were more complicated. The first date mentioned would usually be the date on which the document was written, but not always, while people and places could concern the act of writing the document, but could also represent aspects of the case which was treated in the document. However, at the time of writing of this paper, we have already become better in training software where to find what in the documents.

One of the guiding principles in setting up the ontology was flexibility: the ontology should be easy to extend with new attributes for existing entities, and even with new entities. Our storage application would have to support this flexibility. Furthermore, we needed a solution that would be both platform independent and application independent. Whatever happens to our IT architecture, we want to be able to continue operating the workflow. Also, we want to be able to deliver the metadata to our customers in an application and platform independent way. The solution was found in a previous project we did for a museum in The Netherlands: an entity database built up as a store of semantic triples. Semantic triples are the atomic data entities in the Resource Description Framework (RDF) data model [7]. A semantic triple is a subject-predicate-object expression, e.g. "Act W-756" "has subject" "last will and testament". The advantage is that, on the storage side, you only need to be able to store triples. Design of different record structures for different types of data is not necessary. This means that extension of the data model is perfectly possible without the need for extensive software modifications.

While identifying entities, we are also building up authority files: persons and places encountered are being checked against list of places and persons that were encountered before. We also have built up a list of authorized subjects. If the workflow encounters entities that are, as yet, unknown, a log message is generated, allowing the operator to decide if a new entity is called for, or that it should be linked to existing entities.

We had to build up our own authority files of relevant people and subjects, but for places we made use of exisiting authority files like GeoNames [8].

The way we harvest and match entities was very much inspired by TextRazor, which "offers a complete cloud or self-hosted text analysis infrastructure." It combines "state-of-the-art natural language processing techniques with a comprehensive knowledgebase of real-life facts" [9]. Of course, where TextRazor depends heavily on Wikidata for it's knowledgebase, we had to build up our own knowledgebase, as the people and subjects in our corpus are mostly not in Wikidata. However, the limited scope of our corpus made this perfectly possible.

Assembling the workflow

Figure 3 shows our improved workflow. OCR is now always happening. Depending on the requirements of the project, either the whole text is read, or only text from targeted areas (to extract the metadata needed for the workflow). We use third party OCR engines. We started out with Transkribus to establish if our approach was possible, but now we also use more main stream engines like Tesseract [10]. Apart from the OCR engines, all other software guiding the OCR and processing the results is our own. These software modules are usually further developments of software we developed in the past five years for quality control in (mass) digitization processes.

As you can see in figure 3, an inventory prepared by the data owner is not strictly necessary anymore. With the metadata harvested from the documents, an inventory can be created on the fly. However, if an inventory with sufficient detail is being provided, the quality of the checks in the last step is markedly improved. Documents can, for example be

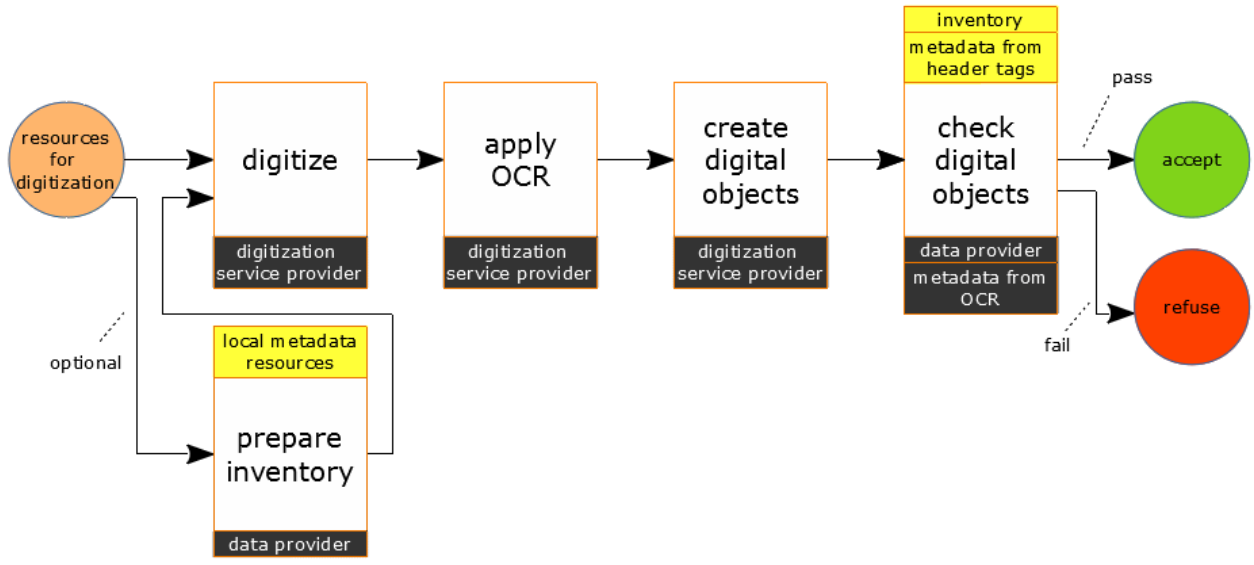

Figure 3. *New mass digitization process*

matched to the inventory on dates and certain key words.

The step “create digital objects” does now involve the attempt to match the entities harvested from the documents with entities found before and, if applicable, the addition of new entities to the database.

The step “check digital objects” now has access to a new resource: the metadata extracted from the actual documents.

While working on the project, we have made the workflows as independent as possible. The less we need to configure, the better. The workflow itself establishes which tools are needed: a printed source - which, of course, we can also deal with – will be treated differently from a handwritten one, and images will be processed in another way again.

While extracting entities, the actual extraction is immediately followed by assessment and validation of those entities. In case of doubt, the workflow sends messages to the operator to ask for confirmation (if a possible match was found) or manual creation of an entity. If the operator confirms a tenuous match, this is fed back into the system.

By now, we are not just extracting descriptive metadata, but we are also automatically generating structural and administrative metadata.

Lessons learned

Looking back, we think we have made significant progress. Combining OCR/HTR and structural analysis of documents, we are now indeed able to automatically extract metadata from a real corpus of handwritten 19th century archival material. Furthermore, we can harvest entities from these documents that allow for automatic classification on date and subject during the workflow with a very low error percentage. Beyond the actual digitization workflow, we are harvesting entities that can be turned into linked data metadata objects that can be delivered to our customers, to allow easy integration of their data in emerging linked data environments.

This does not mean that this is a finished project. Huge progress has been made over the past twenty years with regard to HRT, but it still presents difficulties.

Although we set up this project in less than 6 months, we do realize that this was only possible because of our previous experience with linked data and ontologies, with quality control in digitization and with digital workflows in general. Otherwise, it would have been a much more difficult and much more drawn out project. Also, the scope of the data selected for the project made our life relatively easy: a fairly uniform set of documents from a limited period of time and a limited geographical area, covering a fixed set of subjects in a few clearly established formats. Had we chosen a more diverse se of data, a lot more work would have been needed on all aspects of the project.

We also realized, once again, that a lot of work still needs to be done on the development of both ontologies and knowledgebases, to use the term preferred by TextRazor of entities and their relations. Looking at the limited scope of the corpus selected for this project, it still took a lot of time to build up the knowledge base, and having built it up, there is, as yet, no ‘authority’ we can hand it over to. As long as these conditions continue, it is difficult to lift the linked data of projects like ours out of the proof of concept stage into a full production environment.

Having said that, the original goals of the project were certainly met: our digitization workflow has benefited directly, and without any doubt from the extraction of metadata from the documents.

The future

The project has definitely returned enough to make it worthwhile to continue the work. Several areas for improvement present themselves:

Firstly, further improvements are possible regarding the structural analysis of documents: it should be possible to make the workflow more independent with regard to establishing where in the documents which entities are likely to be found.

Secondly, we want to look further into the word graph approach mentioned before. We are quite happy with the results of the ‘conventional’ OCR and HRT methods we are applying now, but more might be possible.

A third ambition, probably to be realized after and as a result of the two points mentioned before, is the improvement of support for ‘mixed’ collections. Uniform collections are being processed very well, but diverse collections of archival material need more work.

We have also started work on different types of documents. So far, we have focused our attention on text, but in recent projects we have started to try to extract metadata from cartographic and musical resources. As a step for the not so far future, we are also considering some work image analysis (for photo collections) leading to metadata generation.

The issue of finding good ‘homes’ for the ontologies and knowledgebases created during our projects will also need more attention.

Finally, we are working on integration of more overarching metadata standards (i.e. METS, PREMIS and MODS) in our workflow, to allow for production of metadata packages that can easily be ‘fed’ to applications for storage end dissemination.

References

[1] https://en.wikipedia.org/wiki/IDEF0

[2] https://www.lotsimaging.com/

[3] Zhen Zhang. Unconstrained Word Graph Based Keyword Spotting, Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering (2013), https://doi.org/10.2991/iccsee.2013.251

[4] https://www.himanis.org/

[5] Alejandro Héctor Toselli, “HMM word graph based keyword spotting in handwritten document images”, Information Sciences, Volumes 370–371, Pages 497-518 (2016), https://doi.org/10.1016/j.ins.2016.07.063

[6] https://transkribus.eu/Transkribus/#scientist-content

[7] https://www.w3.org/TR/rdf11-primer/#section-triple

[8] http://www.geonames.org/

[9] https://www.textrazor.com/

[10] https://github.com/tesseract-ocr

Author Biography

Martijn van der Kaaij (Amsterdam 1971) is a founding partner of Heron Information Management LLP. As part of his master's degree in history, he studied the application of ICT to the arts and humanities, which developed into an enduring fascination.

Martijn has 21 years experience of delivering training on metadata, process management and work flows. For Heron, he also provides consultancy on these subjects, and develops software for both quality control and dissemination of metadata.