Modern Times (Semi-)Automatic Enrichment of Metadata

Eric de Ruijter

MODERN TIMES (SEMI-)AUTOMATIC ENRICHMENT OF METADATA

Eric DE RUIJTER

Serving our own users

A street in Amsterdam, houses, people walking on the streets, maybe bikes and cars. That's how most people might describe any given photo from the IISH collections. However, these are not the words used in the institutional catalogue. Instead, there, it says this photo is about strikes and tramways (in Amsterdam in 1955).

This is not surprising, the International Institute of Social History, as part of its mission, has been collecting and making available materials on labour, social movements and demonstrations, for over eighty years. A tram not showing up is the reason why this picture is a part of the collection.

Until the end of the 20th century this was what users expected to find at the Institute. This rather limited group of users knew the institute and were looking for books and archives in the realm of social history. And the institute also knew its users, their expectations and questions. This is no longer obvious.

New users and more data

We will have to reconsider our policies; in the first place the users no longer go directly to the institute to find their material. They could start anywhere, with data being distributed through several search engines and platforms, such as Google, Europeana, WorldCat, or Archives Portal Europe. They might never have even heard of the institute, maybe it is not clear what they are looking for, or might be interested in.

At the same time, the amount of digital data is growing, and it's growing fast. Many new collections are being acquired in a digital form. This means that more and more data will be ingested into the digital repository, sometimes without much information about the content of all the files. Handmade descriptions will be brief and limited.

So, to be able to reach out to more diverse user groups and to handle large amounts of diverse data, we have to look at more large-scale and (semi-) automated description methods. Also, the new group of Digital Humanities

researchers can't wait to use the collection data. Although they might want to use their own tools, we have to meet them halfway by opening up our collections. Automatic description and enrichment is no longer a choice, but a necessity, and we have to experiment with it.

From physical collections to data

The first requirement for better, or extra, descriptions is to turn physical collections -as much as possible- into digital format; into machine readable data. Collections have to be considered more in terms of data, that's the basis for deep searching, recognizing entities, or annotating text. The IISH repository holds about 100 terabytes worth of digital collections, both digitized and digitally born, but the fact is that the best part of the 50 kilometres of analogue collections have not been digitized yet. Although in the coming years we will still not come near complete digitization, we have to continue trying.

However, while scanning texts and images, digitizing audio and video is required, it is not enough. The next step is turning these images (or audio) into machine readable text. (ALTO-)OCR is already part of most digitizing processes when it comes to printed material, but many older collections are handwritten. Handwritten Text Recognition (HTR) requires more investments from the beginning, but good results have already been achieved for large text corpora, such as the 18th century resolutions and reports of the Dutch East-Indian Company. In the case of audio and video materials speech and image recognition might come to help in future digitization.

Text mining and linked data

In 2018 the IISH started a project to analyse its encoded archival descriptions. Based on a research project[1] on the recognition and extraction of historical entities from a variety of historical sources, an application is being developed that can recognize entities in EAD descriptions. These persons and organizations can then be more easily found in an institutional catalogue, but it's also an essential step in creating meaningful linked data.

Then the next step is to use the entity recognition tools to analyse born digital archives. As it won't always be possible to make a classic inventory, these techniques will help to extract all kinds of information from the digital archive and give insight into the collection. A future inventory might appear as several visualisations of networks of people, or overviews of important events, timelines and maps.

Creating linked data is an equally important step. For the IISH, the intention was to improve its website by making its data re-usable and interoperable. To do this we aggregated several sources of data: library records, Encoded Archival, audio-

[1] Original HiTime CATCH project: https://ilk.uvt.nl/hitime.

visual descriptions and research data sets. The data was marked up as RDF, saved in a local triple store and made available together on the new IISH general website. Every data source has its own pipeline of extraction, transformation and loading into the triple store, this is a daily automatic routine. Our linked data use more general schemes, like schema.org and Dublin core, and are linked to external vocabularies like VIAF, AAT and LOC subject headings in a combined machine-human action. For the time being we will continue to use traditional metadata management systems. The linked data conversion also helped evaluate our efforts to date, it exposed the weak spots in the source data, such as ambiguous authority records, and enabled us to improve them.

Diamond workers as data

The ANDB archive (the archive of the Dutch diamond workers union) based in Amsterdam is a good example of an analogue collection going through the above-mentioned pipeline, transforming it into linked data. It was digitized as part of the Dutch national programme for the preservation of paper heritage (Metamorfoze). An interesting part of the archive were the membership cards from the beginning of the 20th century. First, the handwritten cards were transcribed through crowd sourcing, in 8 months 27,000 cards were turned into 'raw data'. Notations of addresses, unions and job titles written on the cards were harmonized and the clean data were converted to linked data. They were linked to a reference list of Amsterdam street names in order to show the members on a (historical) map. By also linking the data to the Amsterdam population register it is possible to interpret more about their religious background and careers.

The ultimate goal of the whole project was to use the data from the archive to present new research on the diamond workers in a book and on a website[2]. Questions that have to be answered for this were, for example, whether the workers lived near the diamond factory? What was the composition of the workforce? How did they emancipate or integrate over generations?

Working together

The tools and applications to work with we can't develop alone and we shouldn't try to, there are two developments that will help us with this.

First, cooperation between heritage institutions is important. In the Netherlands, we are privileged to cooperate within the Network for Digital Heritage[3] (NDE), supported by the government, a national strategy to make digital heritage sustainable, usable and visible. It encourages the creation of linked data networks, share tools and stimulates heritage software suppliers to adapt their applications.

---

[2] Website of the Dutch diamond workers: https://diamantbewerkers.nl/en.

[3] NDE, Dutch digital heritage network: https://www.netwerkdigitaalerfgoed.nl/en.

This type of cooperation should be put into practice in more local projects. The IISH took part in a local project to bring together the data on Amsterdam from five heritage institutions in a linked data platform[4]. Already, students are getting more used to working with these data, in this project the data were offered as SPARQL queries to students with the assignment to tell a story about Amsterdam. They didn't know about linked data, but understood soon the principle and were soon able to make their own sparql queries. It resulted in nice applications, one showed the growth of the city or another the distribution of protest actions.

Secondly, help comes from the researchers, especially the digital humanities researchers. In the Netherlands the CLARIAH project[5] is building a distributed infrastructure for the humanities, the IISH, and its partners within the KNAW Humanities Cluster[6], are deeply involved in this project. A growing number of researchers are doing large-scale research on data and are in need of the appropriate tools and data. Although the tools are often for a very specific set of research questions and only temporarily in use to be able to get enough results. From a collection viewpoint an effort should be made to apply these tools more generally and for longer term for all kinds of collections.

New digital archivist

Working with automated tools to describe and enrich digital collections will also have impact on the work of the archivist/librarian. We will definitely need an accessioning archivist to register and take control of new acquisitions. But processing collections will merge more and more with data processing: digitizing, transcribing, cleaning, entity recognition and linking, dependent on what will be useful. A (human) archivist should also accompany these (semi-) automatic processes to be in control of what happens.

Audio transcriptions, for example, can be done automatically. There are already tools which promise an accuracy of over 90%, but it's important to monitor the results. In the IISH collection an interview about a major strike in the Second World War was automatically transcribed and the tool recognised words like Google, internet, Obama. Apparently, a dictionary with words from the last ten years was used.

In short, you need an 'archivist' to understand, monitor, and manage this process. We can't leave it all to machines and software. The archivist should be aware of the shortcomings of these kind of tools and intervene where necessary. The archivist

---

[4] Adamlink, a linked data platform of Amsterdam heritage institutions, http://lab.adamlink.nl.

[5] CLARIAH, a distributed infrastructure for the humanities and social sciences: https://www.clariah.nl/en.

[6] An alliance of three institutes of the Royal Netherlands Academy of Arts and Sciences (KNAW) in which humanities research is carried out using advanced methods: https://huc.knaw.nl.

shouldn't walk around like Charlie Chaplin in Modern Times, trying to fight the dictatorship of the machine, we can work together.

The archivist should approach collections with the mentality to connect these to other internal/external collections, therein lies the challenge. For instance, the IISH received a photo collection together with descriptive metadata, the work was not only to try and convert these descriptions to standard archival metadata but to enrich it. We cleaned the data, tried to structure the concepts as persons, organizations, and places in order to link these with useful vocabularies. In the long run, image recognition might add extra semantics to the photos in the hope that this collection might have a broader use.

To conclude

(Semi-) automatic description and enrichment is the way to go forward. Monitoring and improving these processes will create more and more change in the job of an archivist. The techniques and tools are not perfect yet, but by experimenting we can make them more widely applicable and sustainable. More importantly, these techniques will help to generate greater and better use of the collections – serving both a broader audience and more specific groups of researchers. In the near future the photo from about the 'Amsterdam tramway strike' could be automatically enriched with the name of the street, buildings, a car brand or the type clothes, all of which visually recognised by tools directly from the image.