<u>FUTURS FANTASTIQUES</u>
Vendredi 10 décembre 2021

Grand Auditorium

Marloes Bontje & Mirjam Cuper, AI in relation to GLAMS: an overview of the findings of the EuropeanaTech task force on AI

Marloes Bontje :

So hello everyone.

First of all we are happy, to be here life. We didn't expect that we could be in Paris due to all the COVID stuff, but it's good to be here. Thank you for inviting us. My name is Marloes Bontje. I work for the Netherlands Institute for Sound and Vision. This is Mirjam Cuper. She works for the National Library of the Netherlands. And we are here to present the AI in relation to GLAMs Task Force report. So we are two members of the task force, which was a task force of 15 people in total. And this is a little overview that we're going to do today. First a little intro, then we move on to the survey. Then interviews and key findings, and the future work.

So the AI in relation to GLAMs task force is a Europeana task force, specifically from the European Tech community. And the European Tech community, I take it, you've heard of it, but we focus on strengthening the collaborations between cultural heritage professionals, especially working in the digitalizing cultural heritage. So mostly people working at R&D departments at GLAMs institutions. And, so we support the collaboration between those and we focus on enlarging the knowledge, and also the sharing of the knowledge. And we support the fact that all cultural heritage, especially when in Europe but also out there, outside of Europe, is becoming more accessible, for everyone, for the public.

In this task force we had a goal to do a horizon scanning. It was more of a horizon scanning exercise and we wanted to find out how far AI and machine learning are integrated in GLAMs institutions, and how developed this application of this technology is in these institutions. And so therefore, we built a task force of people from the Europeana Tech community. And so we had a strategy to first conduct a survey. And then move on to in depth interviews. And the outcomes that we were aiming for is that we want to, like I just said, get an overview or more insight on the progress being made in with AI and cultural heritage world and also on the different levels: so the identity, the type of media, or data that is targeted and also the methodologies. And we would like to have a formative basis upon how Europeana can facilitate or accommodate innovation and ethical, sustainable growth of AI technologies, within the cultural heritage sector.

So we had quite a lot of questions to ask to GLAMs institutions. These are the things that we wanted to know:

- The level of interest, so how much expertise is there already, within these organizations

- The team size: how many people are working with these topics within an organization

- Type of data they're working with

- The tools they are using

- Data training

- What phase their research is in, so it's still experimental phase, or production of phase, or does it already have a user interface, so that the public can access their collections

- Ethics of course

- If they do performance evaluations already on the work that they've done

- We ask about the challenges they have encountered and also possibly already sold

- And yeah, the key lessons that were learned and how they proceed in the future.

So we can move on to the survey. I will not tell everything that we've done with the survey but I will pick a couple of things. So the first basic thing is the respondents and what you can see in the chart is that there is an obvious Western European bias and also a lot of Dutch respondents. I think that's because of two reasons: Europeana has its headquarters in the Netherlands, five out of 15 Task Force members are Dutch, so when you start sharing this survey on your network, it probably reaches more Dutch institutes. But this is something we should think of for next time, that we could do maybe also a little bit more effort to reach institutes, like direct emails and really asking a lot of institutes to reply to a survey as this.

Nonetheless we have 56 respondents from 20 countries. We're very happy with those, and we were able to pick interviews from these respondents as well, so that's good. And then we had the question of what level of interest do you take, and on what level of expertise are you at this moment? And so, as you can see behind me, there is an obvious interest in certain topics such as knowledge extraction, the metadata quality, and the visualizing of the GLAM collections, collection management and the discovery in search. And in the side table there is also made visible if that has already been applied, and how that was perceived. So on the last chart (I'm not gonna bore you with all these charts, with just a couple of them, so there's the last one) this shows the project goals in relation to the type of media, the type of data. That was that data researched (and there are something that you can figure from this chart). For instance the institutes that work with images and text based data they are obviously quite interested in enrichment. And the automatic indexing is a necessary topic for audio-visual archives. It makes sense because for instance, if you have a daily ingest for your archives, such as TV archives like the place I work (I work for the basically national TV archive). And if you have a daily ingest, you need to process a lot of data daily and so tools like facial recognition, voice recognition, but also the automatic creation of still images that represent episodes or programs. Those are all things that come in handy if you need to store this type of data. I used to work for an another company that was also processing Dutch TV shows and we used to create the stills of episodes manually. So can you imagine how much work that is, that basically means that you run this show in high quality version on the screen and literally push the print screen button. So at some point we were requesting someone to please build a tool for that. And the tool eventually happened so that's a good thing.

From this survey for several respondents we filtered them, because we also ask people to come with a use case and not everyone was able to bring one. And from the use case entries we asked them if they were interested in doing in depth interview. And not everyone replied, but that those replied, we made a selection and that selection was based on also the amount of time that we had to do these interviews. So the number of eight is a random pick and also the distribution over the type of institutes might not be super equal as you can see. But it was, I think, more important to have a distribution among them the type of data that they are researching so that least have a coverage of people working with images and people working with audio-visual material, people working with text. And I think we got that, so that's a good thing. And of course, these people were able to deliver a use case and talk about projects that were in several phases of their operation.

So one of the in depth interviews here is Mirjam Cuper's story. She will now elaborate on this case study and tell a bit about what challenges she faced as a data scientist at the National Library of the Netherlands.

Mirjam Cuper:

I was indeed one of the people for the in depth interviews so I have offered this use case as an example of projects from the National Library of Netherlands.

Every year we have two researchers (let's go) come to the National Library and work with our data, and this use case is from one of those researcher projects. And in this use case in this project, the researcher wanted to know if we could find out which optical character recognition quality is needed for machine learning tasks that are often used by digital humanities students. So we needed the data for that, and because we wanted to see how good a machine learning task was compared to a ground truth, we needed data from which we have both the ground truth and the original OCR available. And we came with the dataset of newspapers and books and what we did then was we had two machine learning algorithms (topic modeling and document classification), and we run both algorithms on both the ground truth set and the original OCR, and the results from the ground truth set were like a golden standard what we would expect as outcome from the machine learning model, and then we looked at the original OCR and we could compare it and see. For example, the OCR is 80% correct, then the machine learning model still works well, but if it's only 40% then it isn't quite this good; and this can be used for digital humanities students to get a better indication of which data they can use for their research and which data not. And we thought as a National Library of Netherlands we always find it's important to share our knowledge so the outcome from this project were a webinar that was publicly available and we gave a presentation at a DH Benelux and there are also a few blog posts about this project, and furthermore we have the code and the data online available so people can continue working with that.

(Next slide please).

And with this project there were a few challenges and I think every machine learning project has challenges. And the three that we found was first that it was really hard to find suitable data and enough data from which we had both the original OCR and the ground truth available because at the National Library we have tons of data. I think we have over 100 million newspaper pages for example but we only have a small set of ground truth data because ground truth is mainly corrected. So it's very expensive to create, and therefore it's very sparse, so it was hard to find enough data to run this machine learning models on. Another issue was the access to assure for a computer with enough computing power because a lot of machine learning algorithms, especially when working on text, they require a lot of computing power and most of the time when you just have a simple laptop it isn't enough. So we had to search for other ways to use enough computing power instead of just the laptop from the researcher and myself. And the last thing was that it takes a long time to run a machine learning model, especially on text, and more especially when you want to train something on text. And it is not only for this project, but I've also seen it by other projects that we always underestimate time it takes to run a machine learning algorithm. So for ICT challenge, if you come up with a project plan that we take into account that we need to save time for the models to run or to train. So we found we came across all these challenges during the project, so we take them with us for the following project because we now know which challenges there are.

(Next slide please).

Well as Marloes already said we did eight of these in depth interviews with use cases. And well, everyone had their own use case and their own challenges and their own outcomes and stuff like that, but we found that there were nine key findings and recurring themes that were mentioned in almost all the projects.

So we thought it would be nice to give a summary of them and the first key finding was about skills and themes. The institutions we interviewed they said that there is need for a cross departmental collaboration and they also said that this is the most efficient when it is done from the start of the project. So when you start the project, it's best to start with a cross departmental collaboration. But they also mentioned that for example, if you have curators in a project and also AI specialist then most of the time there is a difference in motivation and requirements but also in the language they speak and this can cause problems in the communication. So from the interviews we found out that it is useful to have a frequent translation between different parties to see if everybody's talking about the same things and to make sure the communication is good. And also a lot of time there was mentioned that there was a lack of skills and expertise in internal staff or there is stuff with this the right skills and expertise, but there's only limited manpower.

Everyone talked about data which is obvious and there were mostly two parts: quantity and the reliability. From the quantity perspective, a lot of respondents said that there is a lack of data with relevant annotations and if there is data with annotations they are not always sure if their quality is good. They say, well, we need more data with relevant annotations in order to proceed with this project. And when we look at reliability, there was a huge range of how people feel about the reliability of the data which went from our data is not reliable yet we think our data is good enough. So it was really diverse between the institutions. We also asked them about the tooling they used and some institutions use commercial tools and other use in-house tools. And when we look at the reasoning behind this, for the in-house tools, the institution mentioned it's good to use in-house tools because you can develop the skills in-house, you have more control over the project and also for legal reasons it can be better to use in-house tooling. Institutions that use commercial tools they say, well, we have not enough skills in-house so we outsource it, we have a lack of manpower and a lack of time. So that’s for reasons to use commercial tooling or outsource the project. We also looked at the integration and applicability. Several institutions have already integrated their AI projects into their existing infrastructure, and other institutions haven't done so yet. And they all say it's challenging to use AI projects and put them in the already existing infrastructure. The reasons they mentioned were for example the amount of data that is needed for AI projects, and also because AI projects are mostly in an experimental phase so, it's not ready for prediction yet.

(Next slide please).

And there were also questions about ethics. For example, how is it with facial recognition and personal data? Do we have to work with it? But also that is important to have awareness example for contentious data.

Another thing was the user experience because the outcome from AI models is not always what the user expects, so it's important to keep that in mind to get a meaningful representation.

A lot of institutions also mentioned that it's difficult to convince others in the institutions the advantages of AI and this also makes it challenging to get enough budget and manpower.

For the project results and data sharing now, there were various outcomes of the projects. For example, open source code and rich metadata and things like that and the way of sharing was also various ways. Some institutions only keep it internal while others shared with whole GLAM community and sharing with the whole community can reduce the carbon footprint of machine learning.

And the last key finding was about encouraging AI update. While all institutions agreed that it's important to work on AI projects and that they have a great potential, but we're not there yet, so there has a few things that to be arranged such as skills and resources.

(Next slide, please).

Well, this was a short summary of the report from the Task Force. But the Europeana Tech is not done yet, so we will continue with, for example, supporting the knowledge exchange on AI, seek further collaboration with other initiatives sharing of high quality datasets for which we also have launched a challenge in January 2021, and also providing input to the Europeana Research and Innovation Agenda.

Marloes Bontje:

And moving on to the last slide. That can be very short. Now I want to say thank you to all our colleagues, our task force members. If you would like to read the report that we've made, it can be found on the Europeana pro website: you can download it there and read for yourself. And I would like to thank all of you to listen to us and also for the BnF to have us here and organize this conference.

Thank you.