FUTURS FANTASTIQUES
Vendredi 10 décembre 2021

Grand auditorium

Zdenko Vozar, The beginning of the AI implementation in the Czech Republic in the context of the National Library

I would like to speak about the projects which are beginning to implement AI or machine learning into the infrastructure of the Czech libraries and where the National Library plays a part. In the content I will speak about the two important projects, one which is about the portable research modules for the Czech digital libraries and the second one is more exclusive and it just appends to the Czech web archive. Then I will say something about the projects in common.

So the first project is about the Kramerius + and Feeder, is like a new interface for digital library, especially for DH. As you could see, this is the classical interface of Czech digital library, where all the data from all the institutions which are fair numerisation of something is going on. Here we have also some licenses which we could see. DL4DHT is the name of the project and it's mainly for the data extraction of digital libraries, for digital humanities. There are three partners: Library of Academy of sciences of Czech Republic, National Library and Moravian Library. Technical partner is Inqool, and various DH researchers. Timeline is quite short, and it's financed from the ministerial. And the main objective is to serve and extensively use of high quality data and metadata for the digital libraries in Czech republic and the second one is easily implement of advanced module in as many digital libraries as possible. When we are building this technology we hope almost all of the 43 institutions which operate with Kramerius could use also this module.

Definition of goals. We are building the two parts. One is Kramerius + is actually the enriched databases and all the ingestion structure for the new data. It contains enriching data flows, optimize data store, it has internal data format and conversion definitions in the import and export. There is also some integration with external databases and you could add on services for content analysis and enrichment. Then the second part is Feeder which will feed the data to the users, and there is basically creation of interface, which will be graphically and user nicely. Post nicely for users, also with the REST API which community could use. And then there is also the creation of the export formats which will be on the export. Here are some parts of the things behind the Kramerius +: there is actual connexion to the standard digital library, there are some API endpoints which are using mainly auto XML, they are using also image server and IIIF, and also they are working with the NDK dissemination information package, which is the original package which will be built as the original package from the LTP (Long Term Preservation). There is also progressive synchronization on the daily basis with the digital library and there are interconnected with the databases for metadata mainly from integrated library system Aleph which is an Union catalogue of Czech Republic. There are also the conversion mechanism because main format which we use in the project is XML TEI to reach all the other formats are converted and we are working inside with this. And there is also scheduler we think about the ActiveMQ.

This is interesting for the machine learning part, which is consists because there are the elements which preprocess data on the output for the users. There is possibility to enrich the data with high precision OCR service, because many data were produced just with the OCR engines which are sometimes a little bit older, or maybe not so updated as well. We are prepared to use PERO OCR which is from the project of the Moravian Library use the convolution networks and recurrent networks with the CTC loss function, then there is possible to use Tesseract (we want to use it in the National Library) and any other OCRFeeder actually. Then there is also the UDPipe, the thing which is already containing a lot of different parts for the tokenization, tagging, lemmatisation and dependency parsing, use actually it's built basically on the multilingual BERT: in 2018 it was one of the state of art tools for this thing. It has very good performance as well. Export formats are CSV, JSON, TEI, XML, what the data that our researchers will want. Then there is the third tool, it's the name tag. It's for the name entity recognition. You could see, it works well on the Czech, language, English, German, Spain and Dutch language, works well with flat and nested entities. It's rather complicated when we are going into the technical part of it, because it uses the combination of the different models and works very well with the person location, numbers, age, postcode, time, date, medias also a cultural artefact which he could recognize. And we use here the LINDAT/CLARIN REST API, but it could be also installed on premise. And then in the future, when the module will be ready, it is also ready for expanding the other modules based on machine learning.

There is the interface for the Feeder, I will be call about it here. It's based mainly on the virtual collections which are basically containing the dataset. The virtual collections are normal for the classical digital library: you could see here something about Beekeeping. However, here is about the exports which you could get from the Feeder. However, for the datasets of the Feeder, Feeder will be containing the datasets which will be curated by the professionals of the libraries and also a technical staff. It's realized via functionality of virtual collection, we are already using something which is implemented in classical library. However, with the Kramerius + we extract the data and we enrich them and send them to the researchers. We could do a lot of these datasets on the demand actually. However, there is thing which is sometimes complicated because the original datasets are never too well curated, so there need to be also the assessment of the content of collection of the technical side of the collection, they need to be correction, maybe sometimes redigitization for the uses of the DH community. For now we are just working with the three model datasets. Here are Czech local topographies, Czech esoteric literature and spiritism, and Masonic literature in Czech lands. There will be part which will be open to the users. And when the legislative in Czech Republic changes maybe the other part which is now non-public will be possible to use in the research also, the database and metadata will be publishable.

[next slide]

Here is just to see the interface of Kramerius +, it's basically the back end. You could see there are all the IDs which you need to work with.

[next slide, etc.]

Here is the feeder interface, it's for the faceted search. It is really hard to design a very good interface for the for the such complex use.

Here is the detail on the collection and you could go also on the detail of one page inside the collection of works inside one work and so on.

Here are some statistical images which are based on the facets and content inside the collection or inside your choose (because you could choose not only the collections but the works from out the collections if they will be imported in the Feeder in the Kramerius +).

Here is the export format TEI XML example. You could see here is actually all the data which you will be needed. You could use them, you could use the metadata or you could prepare yourself actually.

Czech web archive advanced data extraction interface

This was like kind of big deal for National Library because they possess Czech web archive from 2000. So now it's like almost 20 years of the operation of the web archive. It has almost 400 terabytes and it contains mainly Czech domain, but also the things which was said Bohemica. It operates on the usual stuff tooling of the web archive. On the project I could say that inside the project is National Library, there is Institute of Sociology of Czech Academy of Sciences and University of West Bohemia Faculty, Department of Cybernetics. It was a 5-year project so we have much more time to do it actually, and the next year we are ending.

Here you could see yearly acquisition of the Czech web archive. Main acquisition was in 2014 when one of the big digitization projects was ending. However, it's mainly around like 23-25 terabytes till 50.

The problematic is that it's impossible really to cataloguize very well the web archive, it contains such a lot of data which are connected very loosely and make traditional cataloguization on this will take many 100 years. You need to comprehend the data of web archive, you need to comprehend how the data was acquired and how the different harvests have the settings for the acquisition. So you need to comprehend the data very well on the minutiae level. You need also to create some safe place for sociological analysis and maybe in future Internet historians. One of the things which we want to do is to run the interactive tasks on this data. And there’s a problem with the deduplication and near similarity thing which you need to solve especially in the [bar view], it's not so common with the books, you know all very well the duplicities but in the web archive is very broad this problem.

Three synchronous goals: there was upgrade of infrastructure philosophy. There was the upgrade of application, interface for people to search the data in web archive and then to define output which is convenient for broader research, so not just for sociologists but also historians and any other professions who are into the web archive data.

There was three phases. First was analytical, one of the important part of the analytical phase was quantitative assessment of the web archive, and also the juridical assessment of the possibility of publish the data of the web archive, because now the legislator in Czech Republic is more tense about the thing. Then was the part of data mining where was developed the BERT models, they were pretrained on the common crawl data because this was a problem which we not expected before. The people in the project could not really very well work with the data from web archive because of the legislature. And there was the four models for text and sound analysis, and for semantic and topic analysis. Then there was set the workflow procedures and the index creation. We are now in implementation phases,

which will be ended with juridical assessment of the possibility of sending the data. What is interesting, we first take the harvest to get to the Hadoop. Then with the Spark and Archive Unleashed tools we are doing the preprocessing however we are not stopping with archive Unleash tools we just take from the data frame output, we use the just text for boiler plate, then we use the BERT analyzers which I talked about and then we are sending the data to Hbase. The Hbase contains everything which was so preprocessed, then the data are indexed for the SOLR1 and user could ask the machine for some data and it's like the re-extractruction for Hbase sending it to SOLR1, and then you could work with his like view of web archive. We are working with the mainly text files but we are processing also sound or anything to .txt and the .txt process further. And the approaches are based on very deep neural networks for document classification, which were developed at the Cathedral of Cybernetics.

Maybe here is not the link which should be… Thank you.

So here I prepared a little presentation of the user interface, it is still on development. However, we need to push it more. You could see here are some harvests and you could see that you could choose your topic filter for the web archive content, which you really want. The topics are pretrained on the catalogue of which we already possess for 10,000 pages. It is not so much in the context of the machine learning, but it performs really well because the catalogue is very precise about the web pages. You could add their URL, you could combine anything actually from the filters and use the logical buttons on the side to create any question or query if you want. You could use also the sentiment and also when you want just to see which data you could get, you could use also the randomizer for the just some view into the web archive.

Here the query was run, Now the data are prepared on the back end. Here you could just choose which kind of output you will want, you could have [URL] text, you could have full text in the future, you could have co-locations and in future there will be also the outputs for the network analysis. I am looking for digitization and Hostivar, Hostivar is our digitization factory. So now the question is running. You could see the processing is going behind, you could see also queries which was before, you could work them, you could name them and then you could store them in your profile.

Okay, I'm now searching for the file. You could see the JSON export of it. And here are the data which was found, you could process in the traditional way if you knew the queries or you could construct your own queries, you could also use REST API to work with output directly. You have the headlines, links, topic, sentiment, plaintext, identification of language, title, URL and identification of the record.

Thank you.

Here are the data export which you have seen already.

Well I want to tell something about the two interfaces, one is mainly for almost all digital libraries which are using a specific software which is very broadly used Czech republic. I think it will be really beneficial for them. It contains also user interface for visual analysis and high quality initial metadata. The second project works with unique raw digital born data. User interface is mainly for the construction of query and some impact analysis. You could see the harvests which are already in your query, and it is also for structuring data into new

metadata to create in the future the index of the web archive which you could import into maybe Union catalogue, or on any other machine which could work with many millions of records. And then in the export: first is images, highly structured text, metadata and technical paradata and the second project, mainly on the extracted text, technical paradata and big volumes of data especially, which was preprocessed for this thing. Together they contain expendable and modular AI, preprocessing layers and tools, predefined collections, own collection management and are easily connected via REST API on different workflows which will be set up by the digital humanists or any other researchers.

Here are the repositories if you will be interested in the projects.

Thank you very much.