Workshop: Automatically extract text, layout and metadata information from XML-files of OCR-ed historical texts.

Mirjam Cuper - mirjam.cuper@kb.nl - @CuperMirjam - 0000-0003-0187-9873
The national library of the Netherlands

Introduction

In the domain of digital humanities, researchers are often interested in analyzing large amounts of historical texts. Most of these texts are digitized with the use of Optical Character Recognition (OCR) software, of which some are manually corrected or enriched. These texts are often stored by digital heritage institutions in a variety of Extensible Markup Language (XML) formats [1][2][3]. To be able to use these texts in most types of analyses, the plain text needs to be extracted from these XML files. XML files can also contain important information regarding the reading order, style, layout information, recognition confidence metrics, and which OCR software was used. Furthermore, XML files can contain metadata about full issues of, for example, newspapers. These metadata files contain information such as the title of the paper, name of the publisher, date of publication, and type of text (e.g. article, advertisement or image). This information can be used by researchers to make specific selections of texts based on these characteristics out of the large amounts of data.

To successfully use the data stored within XML files, specific knowledge is needed. Participants of this workshop will get hands-on experience and guidance while learning how to extract relevant information from these various types of XML formats with the use of Jupyter Notebooks. Furthermore, they learn how to restructure this information and how to process the extracted data for future use.

In this workshop, we will work with the XML formats as commonly provided by the KB, the national library of the Netherlands [4]. We will work with the formats ALTO[5], TEI[6] and PAGE[7] to learn how to extract plain text, metadata and additional information regarding these texts. We will also show how to work with the Didl format, which can be used to extract metadata from newspaper articles. The participants will get access to ready-to-go Python scripts which can be reused after the workshop.

This workshop will address the following questions:

- What are XML files and how are they structured?
- What types of information can be stored in XML files?
- Which Python packages can be used to process XML files?
- How can we select and extract relevant information?
- How can we restructure the information into a more readable format?
- How can we automate this selection and extraction process for batches of XML files?
- How can we store the extracted information into other file formats for future research?

The data for this workshop will be provided by the KB. The data will contain XML files with metadata of newspaper issues, and digitized texts of newspapers and books in various XML formats.

Target audience

The target audience for this workshop are textual scholars, digital humanists and other students or researchers interested in working with textual data stored in XML files. Although the used Jupyter Notebooks are self-explanatory, basic knowledge of Python is a plus. Experience with XML files is not necessary.

Requirements

To be able to attend this workshop, participants need to have an instance of Python 3 and Jupyter Notebooks installed on their laptop. To be able to follow the instructions of the workshop, we advise participants to use Anaconda[8] to install these requirements. Furthermore, the participants need to bring their own laptop to the workshop.

Workshop program

First part – theoretical background and practical introduction

Depending on the skill level of the participants, we will start with a short diversion into Python and Jupyter Notebooks, which will be used in the practical second part of the workshop.

Then, the workshop will concentrate on the theoretical background of XML files. We will explain the rationale behind XML files and how they are structured. We will unravel the XML tree with its root, elements, and attributes using various real-life examples. We will also talk about the importance of namespaces. Furthermore, we will demonstrate which information is stored in the various parts of XML files and how they are stored in the XML tree. This information is important in order to be able to later on correctly extract the relevant information from these files.

Finally we will delve into the different packages that can be used to explore XML files and extract information from them. We will introduce the packages ElementTree and Beautiful Soup. For the packages introduced in this workshop the pro’s and con’s will be explained and illustrated with examples.

The morning session will end with a first hands-on practical session, in which the participants learn how to install the packages and how to work with them using a simple XML file and Jupyter Notebook . We will compare the methods and results of the two packages. Participants will also learn various ways of restructuring the data in Jupyter Notebook and how to store this in different types of formats for future use (e.g. text files or comma seperated files).

Second part - practical session

The afternoon session will start with the exploration of the XML formats Alto, Tei and Page. Through these explorations, the differences and similarities in use and function of different styles of XML files/structures will be explored.

Then, we will dive more deeply into the practical execution of working with the before mentioned XML-files in Jupyter Notebooks. Participants will be guided through the different steps needed for processing these XML-files. We will start with a plenary example on how to handle the various steps. With a few assignments, participants will then learn how the obtained information can be used to extract various types of information from the XML files. These assignments will include extracting plain text and various types of metadata, such as article information and reading order. After the plenary sessions, participants can choose to either follow another hands-on example with the instructor, or to work separately on other assignments.

Since research in digital humanities often relies on large amounts of data, the workshop will conclude with automating the previous steps. This can be used to automatically select and extract information from large batches of XML files,thereby saving a lot of time compared to performing this task manually.

References

[1] https://lab.kb.nl/datasets
[2] https://data.bl.uk/digbks/
[3] https://pro.europeana.eu/page/edm-documentation
[4] www.kb.nl
[5] http://www.loc.gov/standards/alto/
[6] https://tei-c.org/
[7] https://github.com/PRImA-Research-Lab/PAGE-XML
[8] https://www.anaconda.com/