Computer automated collation with CollateX and Python

Ronald Haentjens Dekker

Collation of different versions of a text is an important part of editorial practice. Collation done by hand is a time consuming and meticulous task. Therefore computer automated collation tools have been developed to support the scholar and automate the process. CollateX (see: www.collatex.net) is one such tool. The tool is available as open source software and can be downloaded and used for free. It can handle the collation of multiple witnesses, up to hundreds at a time.

CollateX can be integrated in the editorial workflow in a number of ways. One of them is by using the programming language Python. Python (see: http://www.python.org/) is a dynamically typed programming language that is easy to learn and is rising in popularity among digital humanists (see: http://programminghistorian.org/contents).

In this session the basic concepts of computational alignment (witness, tokens/tokenization, normalization, variation, parallel segments, alignment table) will be explained.

It will be shown how input text files can be fed to the collator using a Python script, how the output of the collation can be transformed and visualized in several ways and how the actual alignment can be manipulated.

The session is meant to encourage the use of scripting languages by scholars and to promote the use of computational approaches in this field of research.