**HAENTJENS — CollateX**

<0 figures>

**Computer-Supported Collation with CollateX**

Haentjens Dekker, R., Andrews, T. L., Birnbaum, D. J., Olsson, L.-J. and van Zundert, J. J.

Comparing witnesses of a text is an important part of scholarly editing. Collation is regarded as one of the scholarly primitives (Unsworth, 2000). Comparing texts by hand can be tedious and prone to error, and it can be made more efficient and reliable with the assistance of computers. This workshop will explain how to use the open-source CollateX1 collation tool to compare witness of a texts automatically, in a way that can be used to produce critical textual editions and other types of comparative documents. Attendees will learn how to prepare source materials in any language (including those that use non-Latin scripts and directionality that is not left-to-right) for collation, how to perform automated collation using CollateX, and how to edit the results.

**Full Contact Information**

Ronald Haentjens Dekker (ronald.dekker@huygens.knaw.nl)

Software Architect and Consultant

Huygens ING

The Netherlands

*Ronald Haentjens Dekker* is a software architect and consultant at the Huygens Institute for the History of The Netherlands (http://www.huygens.knaw.nl/dekker/?lang=en). He has been the lead developer of CollateX since 2007.

Tara L. Andrews (tla@mit.edu)

Assistant Professor of Digital Humanities

University of Bern

*Tara L. Andrews* is assistant professor of digital humanities at the University of Bern. Her research interests include Byzantine history of the middle period (in particular, the 10th to 12th centuries), Armenian history and historiography from the fifth to the 12th centuries, and the application of computational analysis and digital methods to the fields of medieval history and philology.

David J. Birnbaum (djbpitt@gmail.com)

Professor and Chair, Slavic Languages and Literatures

University of Pittsburgh

*David J. Birnbaum* teaches digital humanities at the University of Pittsburgh (http://dh.obdurodon.org) and has been enhancing CollateX to collate medieval Slavic manuscript materials. Links to some of his digital philology and other digital humanities projects are available at http://www.obdurodon.org.

Leif-Jöran Olsson (leifjoran. olsson@svenska.gu.se)

Language Technologist and System Developer

Department of Swedish

University of Gothenburg

*Leif-Jöran Olsson* is a language technologist and system developer at the Språkbanken (the Swedish Language Bank; http://spraakbanken.gu.se/eng/personal/ljo). He is also a developer of the open-source eXistDB XML database system ( http://existdb.org/exist/apps/homepage/index.html) and has worked on a plug-in to integrate eXistdb and CollateX.

Joris J. van Zundert (joris.van.zundert@huygens.knaw.nl)

Researcher and Developer in Computational and Digital Humanities

Huygens ING

The Netherlands

*Joris J. van Zundert* is scientific researcher and developer in the field of digital and computational humanities at the Huygens Institute for the History of The Netherlands. A scholar of medieval Dutch literature by training, his main interest as a researcher and developer lies in the possibilities of computational algorithms for the analysis of literary and historical texts, and the nature and properties of information and data modeling in the humanities.

**Target Audience**

Scholars who are interested in using tools to facilitate humanities research, especially with respect to preparing digital critical editions. Participants who wish to work with their own materials will need to bring them (in plain text or TEI markup); the organizers will provide sample data that can be used by participants who do not have their own project materials. Participants are strongly encouraged to install Python 3 and CollateX in preparation for the workshop; the workshop organizers will provide installation instructions in advance. No prior Python programming experience is required. Based on prior workshop experience, we anticipate attracting between 15 and 30 participants.

**Special Requirements for Technical Support**

A computer projector (HDMI or VGA) will be required for the presentation. Participants will be required to bring their laptops, and the room will need to provide sufficient plug-in electrical connections and wireless Internet connectivity for all participants.

**Intended Length and Format of the Workshop**

Full day, two sessions.

*Session 1. 9:00–12:00: The Basics of Automatic Collation*

The first session will cover the theory of collation, the basics of using CollateX, and the collation of plain text data. No prior experience with collation tools is required.

• Introduction to the theory and uses of collation.

• The collation data model: witnesses, tokens, and tokenization.

• Installing, configuring, and testing CollateX.

• Collating plain text strings and files.

• Output options and postprocessing.

• Introduction to normalization.

*Session 2. 13:00–16:00: Collating XML (including TEI) Data*

The second session will cover more advanced topics, most notably the collation of transcriptions that contain XML (including TEI) markup.

• The collation data model with XML (especially TEI) input.

• Advanced normalization.

• Recognizing and tracking markup information during collation.

• Processing tokens differently according to markup information.

• Output options and post processing for XML (especially TEI) output.

**Call for Participation**

We asked applicants on relevant mailing lists (such as Humanist, TEIL, Digital Medievalist) to tell us about their interests, needs, and prior experience with respect to collation. The instructors listed above will serve as the workshop program committee. For participants, up to 30 participants were to be accepted.

**Note**

1. The main CollateX website is http://collatex.net. CollateX Python is freely available in the Python package repository: https://pypi.python.org/pypi/collatex. The source code is open and available at https://github.com/interedition/collatex. For a report about a recent application of CollateX, see Haentjens (2014).

**References**

**Haentjens Dekker, R., van Hulle, D., Middell, G., Neyt, B. and van Zundert, J.** (2014). Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project. *Digital Scholarship in the Humanities* (2014),http://dsh.oxfordjournals.org/content/early/2014/12/02/llc.fqu007, http://dx.doi.org/10.1093/llc/fqu007.

**Unsworth, J.** (2000). Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This? In *Symposium on Humanities Computing: Formal Methods, Experimental Practice*. London: King’s College, http://people.brandeis.edu/~unsworth/Kings.500/primitives.html.