Classics@18: Nury and Spadini

MARCH 24, 2021 | Classics@

Automatic Collation Tools and Virtual Research Environments

Elisa Nury and Elena Spadini

In this contribution we examine the history of automatic collation, which is the comparison of different copies of a work, to understand how software for this philological task has been (or has not been) integrated within Virtual Research

Environments (VREs). We have identified two general approaches: the first is that of the all-in-one integrated VRE and the second is that of the modular pipeline composed of various independent tools.

First approach: all-in-one VRE

Since Dearing (1962), who is one of the pioneers of automatic collation, and up to the end of the 1990s, collation tools came mostly as part of VREs that were meant to produce a complete and printed scholarly edition, with the established text and its critical apparatus. For instance, Collate (Gilbert 1979:245–246) was designed to produce a critical edition of a medieval prose text.

In practice, these programs read the texts, which were given to the machine through punched cards, magnetic or optical tapes, or disks. Then the program compared the texts and printed out the variants to the base text. These environments might include mechanisms for preparing the texts, neutralizing certain kinds of variants such as orthographical differences or post-processing the results.

In this first approach, the results of the collation were valued as part of the process of *recensio* with the aim of justifying, enriching, and documenting the critical established text of the scholarly edition. Therefore, the collation is instrumental to the creation of the edition, which is the final product. It just happens to be that inside the environments' collation was quite central, as we can see from their names: *Collate*, in the case of Gilbert (1979), *Opcol* in the case of Cannon (1976), and again Collate, the well-known program by Robinson (1989).

For example, Gilbert (1979) describes the different modules of her collation program in details for the edition of Buridanus's *Quaestiones super libros Metaphysicae*. The first module joined together the lines that were separated on several punched cards and tokenized the text into single words. A "strip" parameter served as a normalization mechanism and recorded special symbols to ignore, for instance punctuation marks, to limit the number of insignificant variants in the results. Once every witness had been read, the next module COLLATE2 did the collation: each witness was compared pairwise to one that had been selected as the base text. Then the next modules merged the results of these successive collations so that each line of the base text was accompanied by a critical apparatus. The editor could manually modify the results. The last module would print the text and apparatus in its final form. [1]

Second approach: modular pipeline

More recently we can observe a shift towards a different approach: the tendency to arrange various independent tools into a modular pipeline, helped by the adoption of a common data model to represent variation – the variant graph – and the decision to focus on text alignment only, instead of the creation of a complete printed scholarly edition.

In this approach one would use a program to transcribe and encode the texts of the different witnesses. Then, if necessary, the transcription data would be transformed into a suitable format for the collation tool which collate the texts. The collation results will potentially be transformed again and reused for stemmatological analysis or any other kind of post-processing such as producing different visualizations. [2] The idea of separating the blocks is not new: already in the 1970s the concept of separation of concerns was adopted in computer science and partially in what was computational philology. For instance, Gilbert's program was in fact made of nine modules, while Robinson's Collate was a suite of twenty-five programs (1989:99).

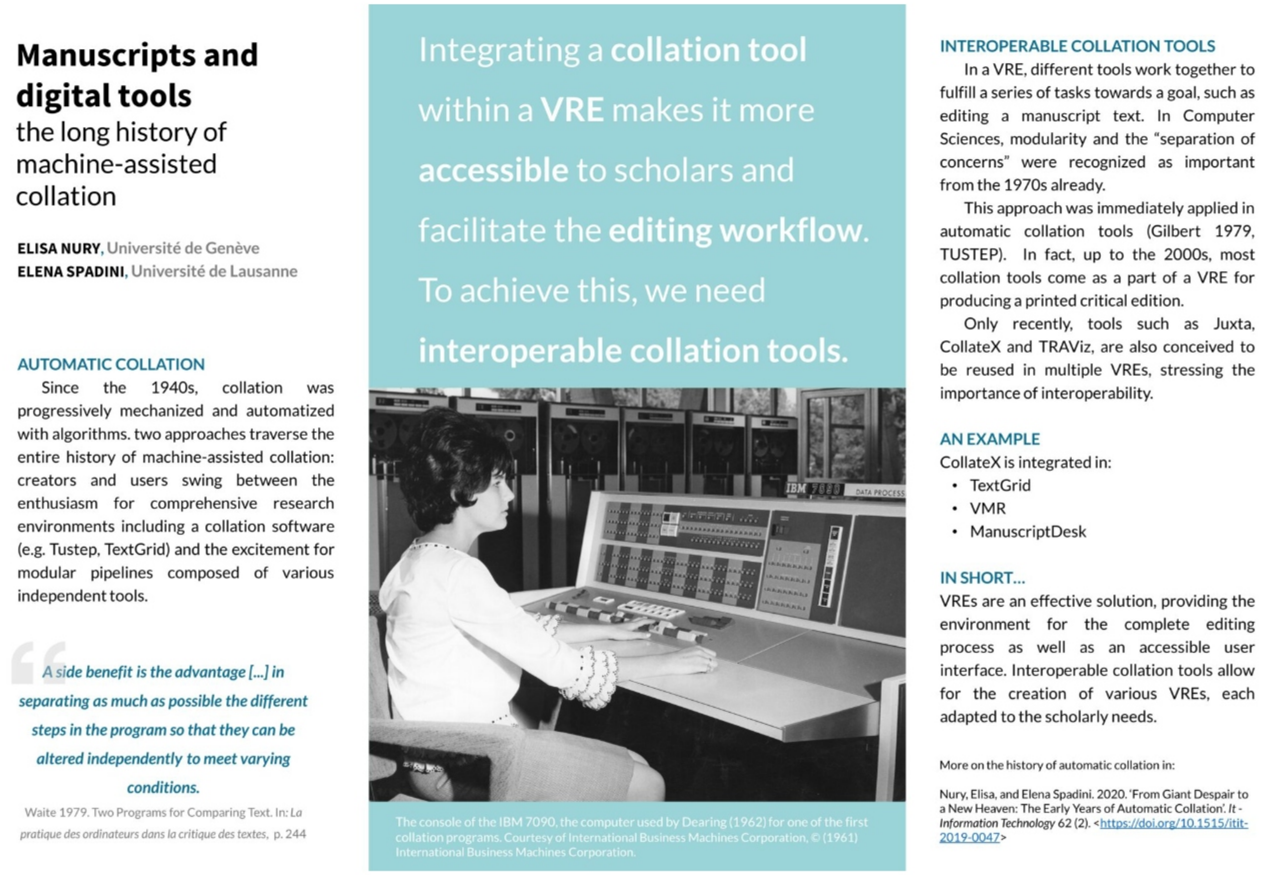
In this second approach, the results of the collation are data, which are valuable per se because they can be exploited, can be computed with multiple aims. For example, one may want to collate witnesses of different periods to study the linguistic changes over time and not directly to produce the critical apparatus of a scholarly edition. Of course, in order for this second approach to be effective the tools that we combine together to form a modular pipeline should be interoperable, which means that they should use standards for data exchange and open APIs for example. [3]

These independent tools can then be integrated in multiple VREs depending on scholarly needs. For instance, CollateX is used in various VREs, such as TextGrid, the New Testament Virtual Manuscript Room, ManuscriptDesk, and a number of smaller projects. TRAViz is another recent collation tool that can be integrated in a VRE: it was adopted for instance by Shor et al. (2021) for the "Tikkoun Sofrim" project.

Conclusion

Both approaches have their own benefits and drawbacks. A modular pipeline may offer a flexible interface fine-tuned for a single project, but it will require more work to code and be less user-friendly.

To conclude, we suggest that VREs are important because they make available to scholars a variety of tools already combined in a seamless workflow: they facilitate the editing workflow, or any other scholarly activity performed on manuscripts. However, to do so, the individual tools should be accessible and interoperable, neither of which is a given. Accessibility firstly refers to the possibility of obtaining the software through a public repository and with a suitable license, whereas interoperability indicates that it can be used in combination with others, for instance by means of multiple open-standard formats for data input and output. Furthermore, the general usability of the tool is relevant for its integration into a larger environment: the quality of its documentation and its ease of installation, for example, are important factors in this calculation. [4] Of a dozen or more collation tools created in the twenty-first century, only a couple have been successfully integrated into a workflow comprising a variety of programs, whereas others such as Juxta stopped being maintained online. Only if the tools are conceived as potential parts of a larger mechanism can we have customized environments to tackle the different problems that scholarly editing poses depending on the materials we are engaging. In this way, we retain the ease of use of the environment and the flexibility of a pipeline adapted to a particular need.



Manuscripts and digital tools
the long history of machine-assisted collation

ELISA NURY, Université de Genève
ELENA SPADINI, Université de Lausanne

AUTOMATIC COLLATION

Since the 1940s, collation was progressively mechanized and automatized with algorithms. two approaches traverse the entire history of machine-assisted collation: creators and users swing between the enthusiasm for comprehensive research environments including a collation software (e.g. Tustep, TextGrid) and the excitement for modular pipelines composed of various independent tools.

> *A side benefit is the advantage [...] in separating as much as possible the different steps in the program so that they can be altered independently to meet varying conditions.*

Waite 1979. Two Programs for Comparing Text. In: *La pratique des ordinateurs dans la critique des textes*, p. 244

Integrating a **collation tool** within a **VRE** makes it more **accessible** to scholars and facilitate the **editing workflow.** To achieve this, we need **interoperable collation tools.**

The console of the IBM 7090, the computer used by Dearing (1962) for one of the first collation programs. Courtesy of International Business Machines Corporation, © (1961) International Business Machines Corporation.

INTEROPERABLE COLLATION TOOLS

In a VRE, different tools work together to fulfill a series of tasks towards a goal, such as editing a manuscript text. In Computer Sciences, modularity and the "separation of concerns" were recognized as important from the 1970s already.

This approach was immediately applied in automatic collation tools (Gilbert 1979, TUSTEP). In fact, up to the 2000s, most collation tools come as a part of a VRE for producing a printed critical edition.

Only recently, tools such as Juxta, CollateX and TRAViz, are also conceived to be reused in multiple VREs, stressing the importance of interoperability.

AN EXAMPLE

CollateX is integrated in:

- TextGrid
- VMR
- ManuscriptDesk

IN SHORT...

VREs are an effective solution, providing the environment for the complete editing process as well as an accessible user interface. Interoperable collation tools allow for the creation of various VREs, each adapted to the scholarly needs.

More on the history of automatic collation in:

Nury, Elisa, and Elena Spadini. 2020. 'From Giant Despair to a New Heaven: The Early Years of Automatic Collation'. *It - Information Technology* 62 (2). <https://doi.org/10.1515/itit-2019-0047>

Bibliography

Camps, Jean-Baptiste, Lucence Ing, and Elena Spadini. 2019. "Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants." Paper presented at the DH2019 Digital Humanities Conference. Utrecht. https://hal.archives-ouvertes.fr/hal-02268348/.

Cannon, Robert L. 1976. "OPCOL: An Optimal Text Collation Algorithm." *Computers and the Humanities* 10(1):33–40.

CLARIAH. 2016. "Guidelines for Software Quality." Task 54.100. https://github.com/CLARIAH/software-quality-guidelines/blob/master/softwareguidelines.pdf.

Dearing, Vinton A. 1962. *Met hods of Textual Editing: A Paper Delivered at a Seminar on Bibliography Held at the Clark Library, 12 May 1962.* Los Angeles.

Gilbert, Penny. 1979. "The Preparation of Prose-Text Editions with the COLLATE System." In *La pratique des ordinateurs dans la critique des textes*, ed. J. Irigoin and G. P. Zarri, 245–254. Paris.

Nury, Elisa, and Elena Spadini. 2020. "From Giant Despair to a New Heaven: The Early Years of Automatic Collation." *IT – Information Technology* 62(2). https://doi.org/10.1515/itit-2019-0047.

Robinson, Peter. 1989. "The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation." *Literary and Linguistic Computing* 4(2):99–105.

Shor, Uri, Vered Raziel-Kretzmer, Moshe Lavee, and Tsvi Kuflik. 2021. “Digital Research Library for Multi-Hierarchical Interrelated Texts: From ‘Tikkoun Sofrim’ Text Production to Text Modeling.” In “Ancient Manuscripts and Virtual Research Environments,” ed. Claire Clivaz and Garrick V. Allen, special issue, *Classics@* 18.

Silva, Georgette, and Harold Love. 1969. “The Identification of Text Variants by Computer.” *Information Storage and Retrieval* 5(3):89–108.

Zenzaro, Simone. 2021. “Towards better VREs: key concepts and basic challenges.” In “Ancient Manuscripts and Virtual Research Environments,” ed. Claire Clivaz and Garrick V. Allen, special issue, *Classics@* 18.

Waite, Stephen. 1979. “Two Programs for Comparing Texts.” In *La pratique des ordinateurs dans la critique des textes*, ed. J. Irigoin and G. P. Zarri, 241–244. Paris.

Footnotes

[back] **1.** See Nury and Spadini 2020 for more details on collation tools between 1960 and 2000.

[back] **2.** For a workflow example, see Camps et al. 2019.

[back] **3.** On key concepts and challenges for VREs, see Zenzaro 2021.

[back] **4.** For a definition of Usability as the sum of Understandibility, Documentation, Learnability, Buildability, Installability, and Performance, see CLARIAH 2016.

CLASSICS@

Issue 18