The « Bibliothèques Virtuelles Humanistes » (Virtual Humanistic Libraries in Tours) : a Collection, or a Corpus?

Marie-Luce DEMONET ***(Centre d'Études Supérieures de la Renaissance, Université François-Rabelais, Tours-CNRS, UMR 6576),***
marie-luce.demonet@univ-tours.fr

1. The goal of the Bibliothèques Virtuelles Humanistes

BVH, or Virtual Humanist Libraries, a digitization project begun in Tours in 2003, http://www.bvh.univ-tours.fr) is to offer two types of digital representations of a selection of books printed during the Renaissance or of manuscripts: the image of a copy (its "facsimile") on the one hand, and its transcription on the other hand, without additions besides corrections or variations that are essential for understanding the text, and TEI encoding. These two goals necessitate the combined efforts of two communities whose objectives, methods and formulations are very different. The difference is further complicated as the elaboration of a corpus for this period presents additional difficulties: librarians and book historians work with "image processing" computer programs, whereas literature and language specialists use linguistic systems (Hyperbase, Weblex, Philologic...). Currently, technical progress has allowed these different approaches to come together, although they have not yet been combined: libraries prefer to offer text-only versions of documents, obtained via OCR or by manual transcription encoded in order to show a readable text alone or along with its facsimile. Linguistic corpora and text databases for works before 1800 are often constituted of modern editions, which are under copyright and impossible to show next to "their" facsimiles—which often do not exist, as they were established from several different reference editions and do not respect the physical presentation. These editions have the obvious advantage of easily lending themselves to searches for data and to detailed encoding. The goal of libraries of text images is entirely different, as research is generally done only on metadata and, at best, on the table of contents which constitutes a minimal indexation, and sometimes is done through a quick round of OCR. In light of these two traditions, and taking into account the reading habits and requests that recent navigational tools have encouraged, the BVH are devoted to conducting research on the two-pronged front of indexing text in image mode, extracting images from images of the pages, classifying and indexing them, and acquiring a significant corpus of transcribed texts. But does the collection of texts constitute a corpus?

2. It is revealing that this proposition of intervention is on the fence between the two initial themes of the congress and we defend the name of "corpus" for the image and text combination, although it was imposed upon us, so as not to frighten financiers and the general public, to call it a "library". We will discuss the difficulties that this grouping engendered, difficulties both theoretical and technical. The digital libraries as they are comprised at the CESR would not be a "corpus", but rather a "collection" as their only commonality is their period of publication, from 1470-1650, identified as the "Renaissance" in the largest sense, including Antique or Medieval texts edited during that period. Even if the collection includes broad categories (classics of the Renaissance - sources of science - legal and political history - philosophy and theology), *a priori* each book on any shelf could fall under one of these categories, all the more as a fifth category, "particular projects,“ allows additions to one or several subcategories (like the "Rabelais" database, the dictionaries, etc.), the only ones in fact that merit the name of "corpus", the rest being composed of standalone books or curiosities. This is how the members of the jury of the National Research Agency (ANR) understood the word "corpus" in France, as it is obvious from

the projects chosen that the notion was relatively vague, even as they excluded a collection of digitized texts in image mode alone. In order for there to be a corpus, there must be text.

3. Going against this analysis, we would defend :

3. 1. The idea that each work chosen corresponds to an analysis of its form and its content. The researchers who are charged with this (under the supervision of Toshinori Uetani, CESR, with the collaboration of Marie-Elisabeth Boutroue, IRHT) examine its interest from the point of view of the history of the book and of the directions of research that obviously reflect the options of researchers at the Center or their colleagues. They hope to render the object upon which they are working accessible to the scientific community, in order to share knowledge, going against the traditional editorial process that consists of offering a "definitive" paper edition once the establishment has finished it. Selection is therefore a tool of anonymous collaboration and results from a step that is one of a researcher, using available funds (as it happens, those of the Région Centre and its partner establishments) that they enrich in so doing: the library of the Museum of Sologne, in Romorantin, possesses a copy of the 1580 edition of Montaigne's *Essais*: although it is not extraordinarily rare, this state of the text merited being offered to the public in order to compare it with that of the BnF's copy (Gallica).

3. 2. The exploitation of image mode, in particular illustrated elements, is specific to the digitized collection: the AGORA program, developed by the computer sciences laboratory of the University of Tours, allows semiautomatic extraction of illustrations, graphs, portraits, typographic material and initials, providing particular indexed and searchable databases(notably thanks to Thesaurus Iconclass). These "processed" works assuredly constitute a corpus which can be subject to precise searches.

4. As for the texts, the standard is to accept the name of "corpus" for a database that gathers together transcriptions of Renaissance texts in French.

4. 1. But what about texts in Latin, Italian, English, Spanish, etc., that are not excluded *a priori*? A multilingual corpus is all the more conceivable as many Renaissance texts contain large fragments of text in Latin (the *Essais*, for example). Even if we initially preferred texts in French, originals or translations, leaving to researchers from other linguistic regions the care of developing their own corpora, the fluidity of TEI encoding allows the possibility of making a "Renaissance" corpus that would not be limited to French and would contain even bilingual or multilingual "alined" corpora. Thanks to the "TEI Renaissance and modern times" application (Tours, July 2008), appropriate encoding of different versions of a text makes a model available that renders the physical description of a text compatible with its logical description (Nicole Dufournaud, CESR, Jean-Daniel Fekete, INRIA). These procedures are taught in a specialized Master's program and during workshops open to students, researchers and librarians.

4. 2. The progress of "Renaissance" OCR (RETRO) also developed in Tours by Jean-Yves Ramel allows the acquisition of text from difficult-to-read printed matter, and it allows correction thanks to form dictionaries compiled from transcriptions and in different languages. Even if acquisition in text mode with an accuracy rate of over 97% still represents a considerable cost for these early printed books (post-correction is always necessary), it allows incrementation of a collection of texts with highly varying written forms, processed by an annotation tool (Analog, developed in Poitiers and at the ENS-LSH by Marie-Hélène Lay) or a "dissimilation" tool (Thierry Vincent, Poitiers); these collections in turn facilitate the acquisition of new texts and allow for linguistic analyses about the uses found with double-checking the text : in the context of course, and in the facsimile. It will also be possible to search the texts for keywords that

constitute the Iconclass thesaurus, offering in this way a selection of topoi present in the texts, in their tables of contents (all transcribed) and in the illustrated elements.

5. In this way, the BVH is a "Renaissance" corpus that is still in the process of being built, a unique but double-headed corpus, one that is dissymmetric: there are obviously many more facsimiles than transcribed texts. Such a corpus is unified by the metadata that harmonize the catalogue of the work and that of the transcribed text thanks to appropriate descriptors (TEI headers, Dublin Core, OAI protocol) and by the search tools that are used on the plain text, but also thanks to the keywords associated with images and with texts, always offering as an anchor the facsimile and the reference to the original work.