The BVH in Tours: digital library of image, text and data

Toshinori Uetani, Guillaume Porte, Sandrine Breuil, Mathieu Duboc

The BVH in Tours: digital library of image, text and data[1]

TEI Conference

Wien, September 30th 2016

Sandrine Breuil, Mathieu Duboc, Guillaume Porte et Toshinori Uetani

1-1 BVH

The Bibliothèques Virtuelles Humanistes (BVH, or Virtual Humanistic Libraries) are a research project run since 2002 at the *Centre d'Études Supérieures de la Renaissance* in Tours, France (CESR, or Center for Advanced Renaissance Studies) in collaboration not only with academic partners like the Humanism Section of the IRHT, the computer science team of the University of Tours or some linguistic laboratories of Tours, Lyons or Poitiers, but also with Libraries and Archives mainly from the French Centre Region.

1-2 Main goal

The objective of the BVH is clearly stated in 2003. It has never changed since: "to develop a digital library of original documents of the Renaissance period, delivering two types of reliable digital representations, facsimile and text, closely linked together"[2]. Fidelity to the original document has always guided our choices. A facsimile contains in principle the entire part of the original document including binding, blank pages, end-papers, edges or back. We describe as precisely as possible its bibliographical and codicological states in order to guarantee the authenticity of its digital version. In the Epistemon corpus, XVIth century French texts are transcribed directly from the original documents. The encoding schema of our specific TEI-Renaissance guidelines is adapted to the spelling and punctuation of the XVIth century French and the page layout of the Early Modern prints. This scrupulous fidelity to the original document makes our TEI files sometimes very complex.

[1] This paper is supervised by Marie-Luce Demonet, founder and former responsible of the BVH research program. As she retired in this September, Chiara Lastraioli, professor of Italian Literature, conducts now the program. Our first English version of this paper is reviewed by Lou Burnard, to whom we express my profound gratitude.

[2] Marie-Luce Demonet and Marie-Elisabeth Boutroue, « Bibliothèques Virtuelles Humanistes de Tours », in *Nouvelles du livre ancien*, n° 112 : « présenter, en étroite interdépendance, une version en mode texte permettant les recherches textuelles à l'aide de moteurs spécialisés et une version en mode image permettant de rester au plus près de l'apparence du livre ancien » (p. …).

Furthermore, we want the "two types of reliable digital representation" to be linked closely together, while digitization and elaboration of electronic texts usually follow two distinct workflows:

- For facsimile, we digitize each page to get the representation of page-image in "pixels";
- For electronic edition, we transcribe pages of text, manually or by OCR tools, to obtain a digital text, with encoded characters in a given format.

The BVH website began releasing its first facsimiles in 2003, and thereafter, it included the Epistemon corpus of digital editions in 2007. Meanwhile, we started a collaboration with the computing team of Jean-Yves Ramel at the University of Tours to improve an OCR[3] for Early Modern imprints. This tool is not yet available, but software has been developed for automatic segmentation, extraction and indexation of image-blocs. For more than ten years, AGORA and RETRO have been used by the BVH team to extract and index typographical features such as ornamental letters, portraits, printers' devices, etc. A specific database of Renaissance typographical material "BaTyR", is developed and managed by Rémi Jimenes[4].

1-3 Two workflows

During our first efforts to adapt the TEI guidelines to encode French Renaissance texts, with the very precious help of Jean-Daniel Fekete, of the INRIA, Nicole Dufournaud and Lou Burnard, we pointed out two important aspects of our datasets:

- First, there is redundancy of data and metadata between two workflows;
- Second, however, this redundancy could make it possible to link tightly these parallel workflows by their metadata.

[3] Jean-Yves Ramel, LI, RFAI is conducting the PaRADIIT Project PaRADIIT (Pattern Redundancy Analysis for Document Image Indexation & Transcription): https://sites.google.com/site/paradiitproject/home
[4] BaTyR presentation (french) : http://www.bvh.univ-tours.fr/materiel_typo.asp

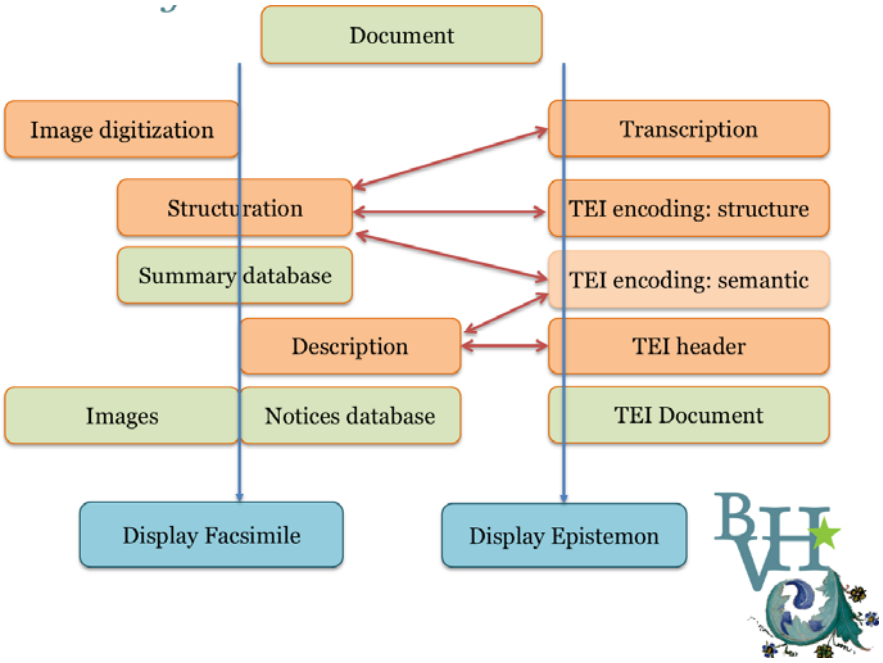

Figure 1 - BVH Workflows: Interconnections between Facsimile and Text

From that time on, our main concern was to merge these two workflows, using a single XML schema and to establish a close correspondence between these two datasets, image and text, incorporating rich metadata[5]. To realize this synthetic vision of Renaissance documents in a digital library, we have tested several indexing-engines and viewers, like XTF (by Berkeley), a framework we are using today mainly to display the textual corpus Epistemon, and TXM, or Mirador viewer with the IIIF protocol, adapted by Biblissima. We have not yet achieved an expected outcome of our representation model.

1-4 "Bordeaux Copy" of the *Essais* of Montaigne

For the study of French Renaissance masterpieces like the *Essais* of Montaigne, the author's handwritten additions and corrections on the printed copy of the 1588 edition were transcribed and encoded by Elise Gautier and, mainly, by Mathieu Duboc under the supervision of Marie-Luce Demonet and Alain Legros[6]. Here are the Bordeaux copy of the Essays of Montaigne (Bordeaux, City library, now available on Gallica, BnF) and an example of encoding of the author's manuscript additions. Example of analysis by TXM engine, a textometric software developed by Serge Heiden and Alelxei Lavrentiev of the CACTUS Team of the IHRIM laboratory in Lyon, for quantifying punctuation corrections by Montaigne himself.

[5] BVH, Dossier Succeed Award 2014 : https://bvh.hypotheses.org/2930
[6] Marie-Luce Demonet & Alain Legros, *Essais 1588 et exemplaire de Bordeaux*, 2015 : https://montaigne.univ-tours.fr/essais-1588-exemplaire-bordeaux/

"Bibliotheques françoises" project

2-1 Project

The "Bibliotheques françoises" project is an attempt to invent a new manner of managing the textual resources to make them available as accurate datasets. In this project, the main goal is to develop biographical and bibliographical databases from an electronic text of two distinct "Bibliotheques françoises" by François Grudé, sieur de La Croix du Maine and Antoine Du Verdier[7]. With a funding from the French infrastructure Equipex Biblissima, Guillaume Porte worked on this project from November 2015 to April 2016 with two student trainees[8]. Furthermore, Mark Greengrass, professor emeritus at Sheffield University, participate actively with this project. His bibliographical database is a decisive factor for our methodological choice. For technological aspects, we are testing the XML-Mind, (XML Editor) XXE, and a XML database, BaseX, in collaboration with Pierre-Yves Buard of the MRSH in Caen and with the financial help of the French consortium CAHIER.

During the mid XVIth century, a hundred years after the invention of printing by Gutenberg, the quantity of information increased considerably and Renaissance men needed a new tool to manage quantity of books and data never seen before[9]. In this context, after the publication in 1545 of the *Bibliotheca universalis* by Conrad Gesner, both « bibliotheques françoises » constitute the first printed dictionaries of French authors. *Le premier volume de la bibliotheque*, – the first volume of the library (but the second was never published) – of La Croix du Maine was published in Paris in 1584, *La Bibliotheque* of Du Verdier in Lyons in 1585. These simultaneous publications testify an apparent competition between the two men[10].

The 2194 prosopographical records of exclusively French authors for La Croix du Maine and about 2300, 20 % of them non-French, for Du Verdier are sorted in alphabetical order of author's first name. According to Mark Greengrass, there are in total 7256 bibliographical records in these two "Bibliotheques". In fact, the two *Bibliotheques* are invaluable sources for

[7] Guillaume Porte & Toshinori Uetani, *Les « Bibliothèques françoises » de La Croix du Maine et Antoine du Verdier (CESR-Biblissima, novembre 2015-avril 2016)*, 2016 : http://bvh.hypotheses.org/2294

[8] Lucas Leprêtre from the ENSSIB and Bruno Farinelli from the University of Torino.

[9] Ann Blair, *Too much to know: Managing scholarly information before the Modern Age*, New Haven and London, Yale University Press, 2010.

[10] Marie-Luce Demonet, Rémi Jimenes and Toshinori Uetani, in *De l'argile au nuage : une archéologie des catalogues (II^{e} millénaire av. J.-C. – XXI^{e} siècle)*, Paris et Genève ; Bibliothèque Mazarine ; Édition des Cendres ; Bibliothèque de Genève, 2015, p. 214-227.

historians not only of literature but also of social networks. We discover in these pages the Republic of the Letters of the late XVIth century France.

2-2 Encoding principles

At this first stage of the project, the physical structure of the whole text of La Croix du Maine was encoded in XML-TEI. Afterwards, we began to mark up inside the <body>, main part containing more than 2 000 records. The <front> and <back> of La Croix du Maine and the whole text of Du Verdier will be processed later. For both *Bibliotheques*, the main element <body>, is divided into 25 chapters corresponding to the 23 letters of the alphabet and two "addenda" division by a with attribute @type="partie".

In each record dedicated to an author or an acquaintance of his, La Croix du Maine often presents some biographical information at first, and describes books written by the latter in a phrase like: "Il a escrit…(some title) / or Elle a escrit…" / "He wrote, or She, wrote…" or "Il a traduit…(title)" / "He translated…" These bibliographical descriptions sometimes contain various titles in one paragraph. La Croix du Maine inserts commentaries, impressions or simple anecdotes. Sometimes, in a single record, we find information about different persons.
In modern dictionaries, each record can be marked up with <entry>, or in a structured list each title can be tagged with <biblStuct>, but these elements don't fit the loose and irregular structure of the two *Bibliotheques*. This is why we mark up each record using again with attribute @type="notice", for "record". This constitutes the basic unit of the prosopographical dictionary. In this *Bibliotheque*, the information about one book can be scattered across several paragraphs. But, if we want to group together different components of the information about an individual person or a single book, we can't use the usual tag <p> for paragraph. For that reason, we mark up each paragraph with the flexible <milestone>[11]. In this choice, as we want to keep a flexible structure, we don't use the element <person>, <occupation>, nor <birth/death/floruit>, but encode these details of prosopographical information using reference strings, <rs>, with an attribute type, like <rs type="occupation">.

2-3 Two working hypotheses

As for the bibliographical references, we imagined two working hypotheses:

[11] Lou Burnard, *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*, OpenEdition Press, 2014 (available online: http://books.openedition.org/oep/679).

The first option is to mark up all relevant information in a <bibl> element, using existing TEI elements like <Author>, <title>, <publisher>, <pubPlace>… and to extract them to constitute a database. With this option, we can establish accurate information directly from the original document, though it costs time, work and money.

The second option is to encode with <bibl> the whole portion of text concerning one or several bibliographical reference(s), even if it comprises several paragraphs tagged with as many <milestone> elements, and to link it to the corresponding record of the specific database developed apart. This solution needs les effort for encoding, but the construction of a specific database needs much more time and money.

In November 2015, we began to encode some examples tagging all relevant bibliographical elements to evaluate cost and time for the first option, when Mark Greengrass suggested his bibliographical database to us in December. From February to April, two student trainees marked up the first three chapters, letters A, B and C, mainly with elements <bibl>, <title> and <persName>. They implemented the xml id of each element <bibl> in the Greengrass's database and xml id of each prosopographical record in a specific prosopographical database; then each biographical or bibliographical data is aligned to online authority files such as VIAF for persons and USTC for printed books. This task of alignment have been facilitated with a very efficient help of Edouard Frunzeanu of Biblissima. It's clear now that, without the database of Mark Greengrass, this first part of the project would have never been completed in this delay.

2-4 Output

This week, we have just released a beta version of the website *Bibliotheques françoises*, developed by Guillaume Porte. It will offer various possibilities of reading and query. Here is its production schema (Fig. 2). This site presents the texts of the *Bibliotheques* of La Croix du Maine and, later, of Du Verdier, prosopographical or bibliographical data in three different modes:

- Records : each prosopographical record from La Croix du Maine and de Du Verdier will be displayed, if a person is in both *Bibliotheque*, the two records, side by side ;
- Prosopographical files with links to authority files (Viaf, IdRef, Bnf…)
- Bibliographic files with links to authority files (USTC, ISTC, VD 16, Edit 16, GLN 15-16) and to digital copies (BVH, Gallica, etc.)

With this website, you will able to navigate among a mass of information collected by the two bibliographers of the late XVIth century. But we'll develop different manners of statistics, visualization of these biographical and bibliographical data. In spite of the title of

"Bibliotheque Françoise", the networks of the persons described by La Croix du Maine is concentrated around several cities, Le Mans, Angers, Tours or Paris. The visualization of their locations will demonstrate clearly this network[12].

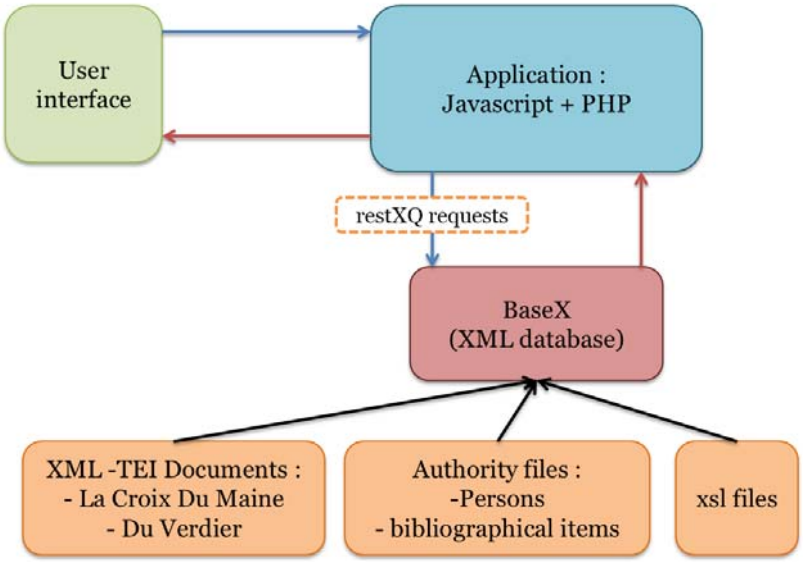

Figure 2 - *Bibliotheques françoises* website: Global operation

Conclusion

The *Bibliotheques* of La Croix du Maine and Du Verdier are principal sources for the literary and social history of the XVIth century France. Nevertheless, their information was not fully exploited, nor was the reliability of their content thoroughly examined. The "Bibliotheques françoises" database provides historians and literary scholars with a useful tool not only to find relevant information from the original text of two bibliographers, but also to re-evaluate the exactitude of each information by comparing it with controlled online authority files and also with digital copies of the original documents. Thus, it contributes not only to improve the data accuracy in the use of these *Bibliotheques*, but also to reevaluate the major current of the French and European Early Modern history. The electronic edition and the database of the "Bibliotheques françoises" project can be integrated in the global workflow of the Bibliothèques Virtuelles Humanistes. Here is its place in the workflow schema designed by Sandrine Breuil (Fig. 3). It characterizes now the BVH as a digital libray of image, text and data.

[12] « Bibliothèques Humanistes Ligériennes » projet partenarial Biblissima, 2016 : https://bvh.hypotheses.org/2665

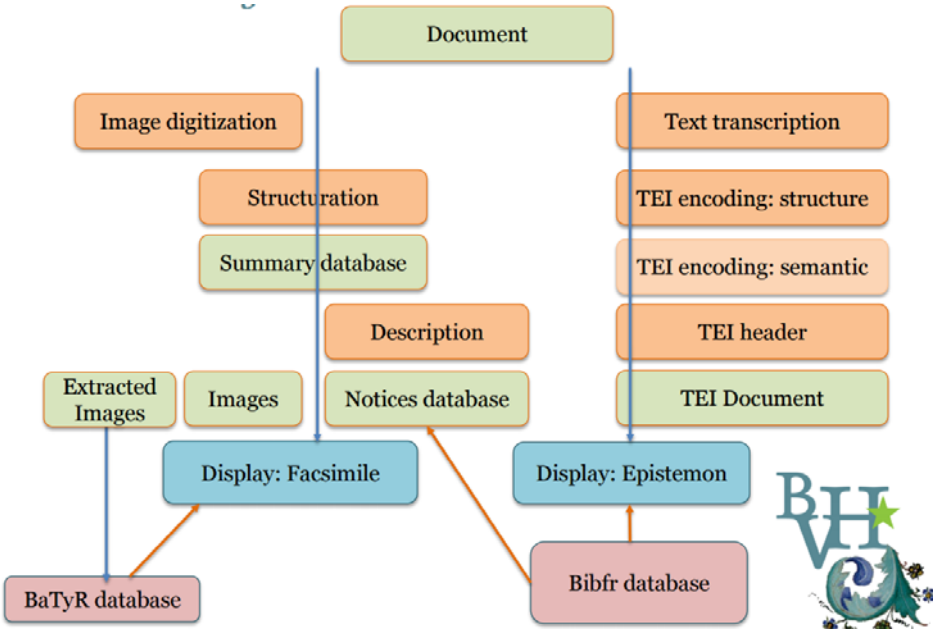

Figure 3 - BVH worflows: Linked by enriched metadata

The problematics and the methodology developed in this project will be shared by young European researchers at the Summer School we are organizing in July 2017 in France, with the funding of the Dariah project: “Humanities at scale” and in collaboration with the City Libraries of Le Mans and Angers, the Library of the Prytanée militaire, ancient Jesuit college, where the young Descartes studied, University of Le Mans and the Ecole nationale des Chartes. We hope for the participation of many TEI members. Thank you very much for your attention.