A Framework for TEI-Based Scholarly Text Editions

Sebastian Pape
University of Kassel
Dept. for Electrical Engineering and Computer Science
D-34109 Kassel, Germany
pape@db.informatik.uni-kassel.de

Christof Schöch
University of Kassel
Dept. of Modern Languages and Literatures
D-34109 Kassel, Germany
schoech@uni-kassel.de

Lutz Wegner
University of Kassel
Dept. for Electrical Engineering and Computer Science
D-34109 Kassel, Germany
wegner@db.informatik.uni-kassel.de

ABSTRACT

In this paper, we describe an interdisciplinary project in which visualization techniques were developed for and applied to scholarly work from literary studies. The aim was to bring Christof Schöch's electronic edition of Bérardier de Bataut's *Essai sur le récit* (1776) to the web. This edition is based on the Text Encoding Initiative's XML-based encoding scheme (TEI P5, subset TEI-Lite). This now de facto standard applies to machine-readable texts used chiefly in the humanities and social sciences. The intention of this edition is to make the edited text freely available on the web, to allow for alternative text views (here original and modern/corrected text), to ensure reader-friendly annotation and navigation, to permit on-line collaboration in encoding and annotation as well as user comments, all in an open source, generically usable, lightweight package. These aims were attained by relying on a GPL-based, public domain CMS (Drupal) and combining it with XSL-Stylesheets and Java Script.

Keywords

Design, Human Factors, Standardization, TEI, scholarly text edition, XSL, CMS, Drupal.

1. BRINGING BÉRARDIER DE BATAUT'S *ESSAI SUR LE RÉCIT* TO THE WEB

This paper is about the framework developed for Christof Schöch's electronic edition [8] of Bérardier de Bataut's *Essai sur le récit, ou Entretiens sur la manière de raconter* first published in 1776 [2], a book which has only recently received renewed attention in French literary studies, and of which only a dozen original copies (cf. Figure 1) are accessible in libraries worldwide. The electronic edition, the coding of which is currently almost completed, allows for alternative text views with user-friendly annotations and navigation. The edition uses the Text Encoding Initiative's XML-based encoding scheme (TEI P5, subset TEI-Lite) which is by now the de facto standard applied to machine-readable texts used chiefly in the humanities and social sciences.

The TEI Guidelines [11] were developed in particular for use in scholarly editions and much has been published on their development and considerable success (see e.g. [9] and the three articles of Thomas Rommel, Allen H. Renear, and Martha Nell Smith therein). For the purpose of our presentation we concentrate on three main goals for serious electronic editions brought forward by one of the TEI-fathers, C. M. Sperberg-McQueen [10]: *accessibility*, *longevity*, and *intellectual integrity*. He argues that one consequence from these requirements is not to tie scholarly editions around a particular piece of software.

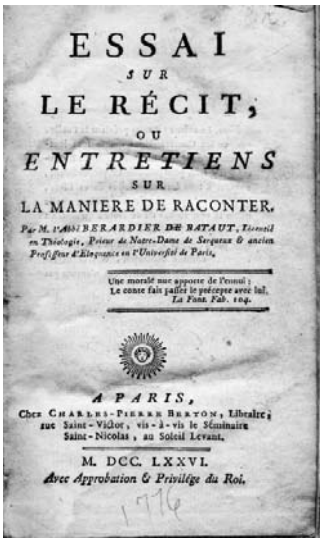
ESSAI
SUR
LE RÉCIT,
OU
ENTRETIENS
SUR
LA MANIERE DE RACONTER.
A PARIS,
M. DCC. LXXVI.
Avec Approbation & Privilége du Roi.

Figure 1. Front page

At the same time, however, he calls for contemporary forms of distribution and use of these archival materials, including appropriate rendering of the *apparatus criticus* for literary or historical research, possibly in different styles and varying degrees of detail for either the advanced specialist or use in an undergraduate class. Indeed, very elaborate examples of electronic versions of scholarly editions exist, more than can be listed here, in various styles and quality levels [12].

One particularly impressive example is the edition of van Gogh's letters [5], a joint project between the Van Gogh Museum Amsterdam and the Huygens Institute – KNAW, which started in 1994 and involved dozens of researchers, editors and technical staff. This edition also uses the TEI encoding scheme and is similar to our framework in that it consists of three main components: (1) subdivisions of the facsimile images, (2) the program that handles the queries to the index, and (3) the rest: static html pages, javascript files, stylesheets, and image files. Because access is much higher than in our case, the workload is distributed over three dedicated servers. The implementation language is Ruby, other tools include the Lucene search engine, the GSV Viewer (now PanoJS), and the ImageMagick set of image preparation tools.

Other noteworthy examples (biased towards French editions) are the Partonopeus de Blois Project of the University of Sheffield [4], the Bernard Barbiche edition of the Édit de Nantes and its predecessors from the period 1562–1598 [1], and finally the online edition of the work of the French naturalist Georges-Louis Leclerc, Comte de Buffon (1707–1788) [6].

Our contribution here is a light-weight framework developed from public domain tools, namely the open source CMS Drupal [3], a new, generic XSLT stylesheet for handling the TEI mark-up, and JavaScript for navigational purposes. The following sections will illustrate each of these aspects with screenshots and explanations of the employed technology.

We leave it to the reader to decide whether our minimalistic, open source approach guarantees accessibility and longevity. Given the vast number of XML-based pages in the Web, we assume that with standard stylesheet and scripting tech-

niques known to work in all current browsers, we are on the safe side.

The third point, intellectual integrity, is not really a technical issue, but our implementation certainly offers everything needed to provide a high level of transparency and explicitness of the editing process. We would like to argue, in any case, that the traditional scholarly edition, now in electronic form with convenient switching between linear transcription and the regularized reading text, contributes more to the preservation and understanding of our cultural heritage than making large amounts of uncommented page scans of historical printings available online.

2. FEATURED FUNCTIONALITIES

When starting an electronic edition, certain basic decisions have to be made concerning the overall organization. One of them concerns pagination. In the case of our Bérardier-Edition each chapter becomes a Web-page which in turn maps into one TEI-encoded XML-file. This is a usability decision whereby the reader scrolls through a chapter rather than clicks from page to page. Annotations which were footnotes then move into the margin which suits well the trend towards 16:9 or 16:10 displays.

The pages from the historical printed version (TEI-element <pb>) become anchors in the text visualized by **[p.*nn*]** as can be seen in Figure 2 below. Navigation to a particular page is possible across chapters as a JavaScript function maintains a chapter-object which knows the start and end page of each chapter and can map target pages to XHTML-anchors inside the proper chapter file. We will not elaborate on this structuring any further, other paginations are clearly possible. Figure 2 also shows the interface for page navigation within the global appearance of the site.

The next basic decision concerns the visualization of the text-critical variants (TEI-element <choice>), in our case the *reg*/*corr*-Version (regularized, texte de lecture) and the *orig*/*sic*-Version (original, transcription linéaire). One option would be a parallel display with synchronous scrolling of both texts. This would be the method of choice in an edition which includes translations (see again [5] as an excellent example). Here, variants concern mostly spellings and single errors either detected by the original publisher as evidenced by the included errata, or corrections and regularizations limited mostly to individual words and typography suggested by the scholar responsible for the current electronic edition.

For this type of edition, a one-column presentation (with margins as mentioned above) with the option to toggle between alternating reading modes is best suited. This permits a larger reading area for one text which in turn gives rise to eye-friendly font sizes and matching spacing. Passages subject to variation are marked with the proper tags in the encoded text and are visually highlighted, in our case with a light-blue background. Mode switching is offered within the function bar at the bottom (see Figure 3 and Figure 4). Details concerning the implementation of mode changes are given in Section 3.

A third major design decision concerns notes (TEI-element <note>). Scholarly text editions distinguish between notes from the author which already appeared in the original source and notes added by the new editor. Here the former are indicated by an asterisk, the latter by a superscripted number running up from 1 chapterwise. Each superscript is placed on a colored background and changes the appearance of the cursor on a MouseOver-event, just like a link.

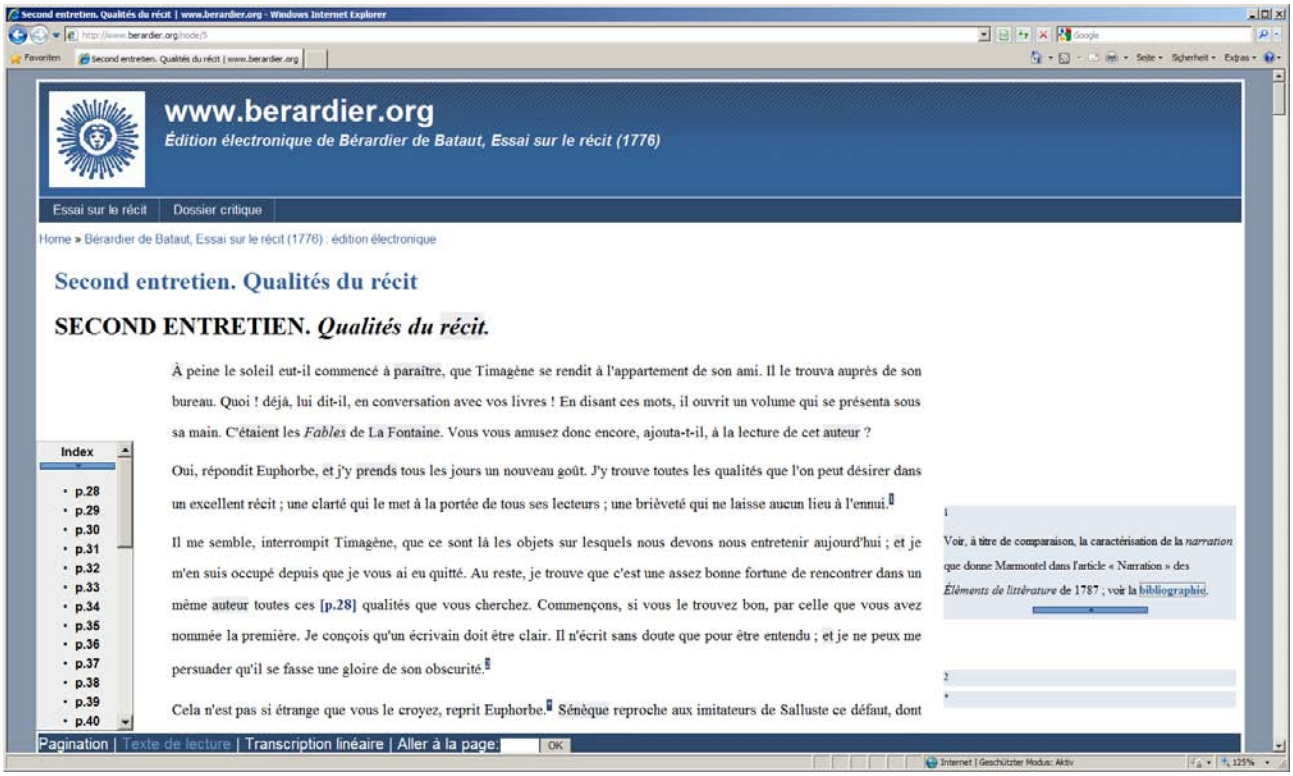

Figure 2. Web-page with function bar including navigational aids

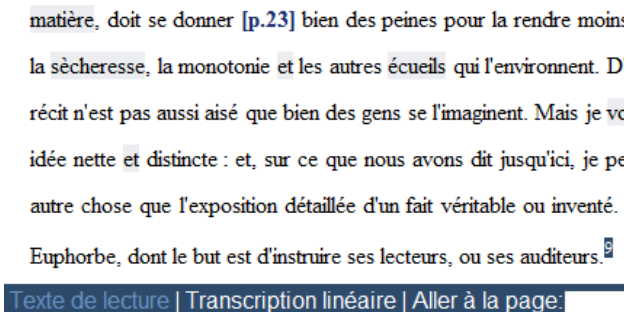

Figure 3. Presentation of modern text version

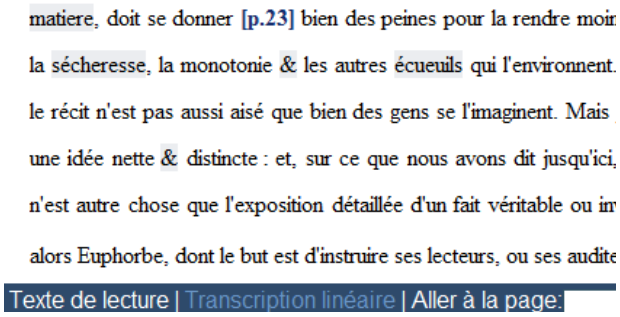

Figure 4. Presentation of original text version

Additionally there is a colored horizontal bar in the margin (cf. Figure 5) on about the same line height with the superscript (numeric or asterisk). The annotations themselves are collapsed in order not to break the flow of reading and only unfold on demand within the margin when the superscript (link) is clicked. The unfolded note then includes a bar at the bottom, which when clicked, closes the note again (cf. Figure 6).

This functionality relies on JavaScript. Should a user choose to turn JavaScript off, the browser will still show the main text but without notes. If several notes are open and overlap each other, the one clicked last is raised, all others are pushed down through their z-index value.

3. IMPLEMENTATION

In this short paper we can only hint at certain implementation details. Since our framework is public domain, the reader is invited to visit the sources which are placed into the documentation section of our project page [8].

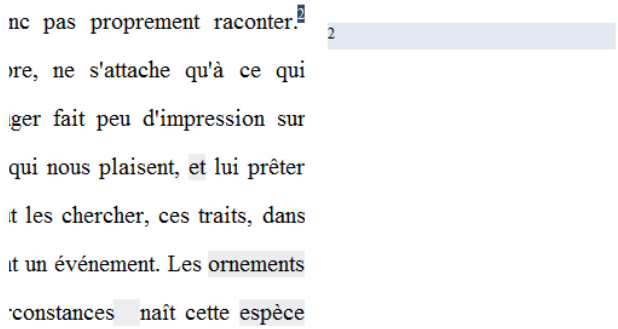

Figure 5. Annotation folded with placeholder in margin

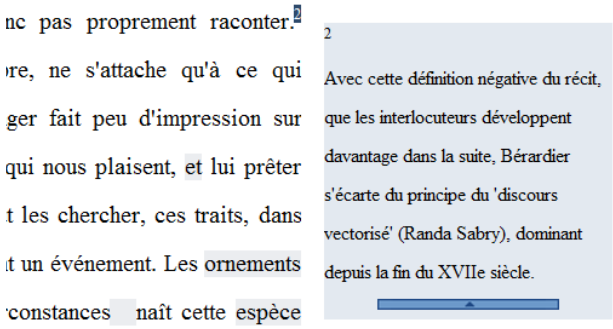

Figure 6. Annotation unfolded with closing bar at bottom

One measure of the complexity (or simplicity) of our implementation is the size of the XSLT-stylesheet [13] which can handle the subset of the TEI-Lite encoding needed for the *Essai sur le récit*. Its size is 25 KB including all comments and has only 20 templates to match. One such rule for dealing with four renderings of <quote> (inline with quotes or without quotes but italic, seperate as block or verse) is shown below.

```
<xsl:template match="quote">
  <xsl:choose>
    <!-- rend="inline" => Quote remains
inline but with french quot. marks -->
    <xsl:when test="@rend='inline'">
      <xsl:text>« </xsl:text>
      <xsl:apply-templates />
      <xsl:text> »</xsl:text>
    </xsl:when>
    <!-- rend="block"  => Quote becomes
seperate block, smaller font, indented,
not italic, no quot. marks -->
    <xsl:when test="@rend='block'">
      <span class="block">
        <xsl:apply-templates />
      </span>
    </xsl:when>
    <xsl:when test="@rend='verse'">
      <span class="verse">
        <xsl:apply-templates />
      </span>
    </xsl:when>
    <xsl:otherwise>
    <!-- rend="italic" => Quote stays in-
line without quot. marks but italic -->
      <span style="font-style: italic">
        <xsl:apply-templates />
      </span>
    </xsl:otherwise>
  </xsl:choose>
</xsl:template>
```

Not all templates are so straightforward. Some insert JavaScript code when translating into XHTML, e.g. when handling annotations. Other templates produce span-Tags which in turn rely on the existence of additional JavaScript files to be called when required by user interaction. Switching between modes as required by the TEI-markup <choice>, calling a JavaScript

function toggleView(class), is one such example, page navigation is another.

For the reading mode switches, the respective XML-elements are grouped by the XSL-Stylesheet into CSS-classes (see Figure 7). The JavaScript function then starts out by collecting all relevant XHTML-elements of the class choice. Next, by accessing the JavaScript object style, it dynamically toggles the stylesheet attribute display of all relevant elements and thus turns the visibility on or off. As not all browsers, namely the Internet Explorer, support collecting elements by class name, the respective function is also part of the JavaScript resources. However, switching modes happens without noticeable delay.

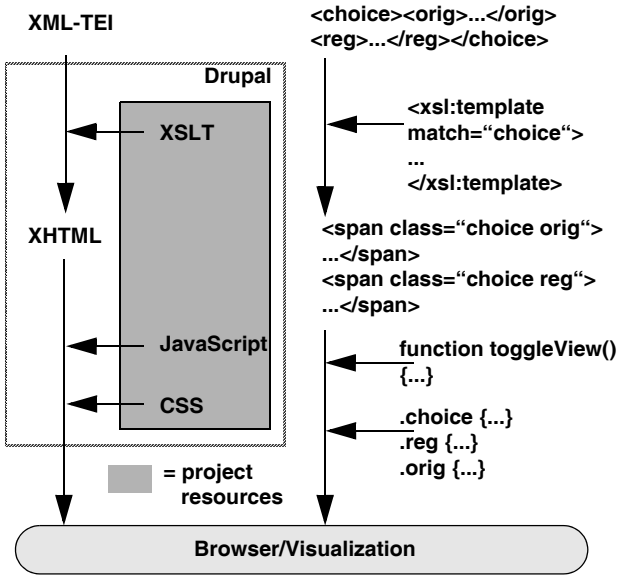

Figure 7. The transformation process

The stylesheet must of course be included into the CMS in order to perform the transformations shown above. Which CMS is used is open to the designer/producer. Setting up Drupal, the CMS chosen here, so that it correctly deals with XML and XSL, requires the following steps: instantiate the database, install the xmlcontent- and book-modules, generate a new input format, e.g. "XML article", choose the XML/XSLT filter, set the path to the stylesheet, select a format under "create content", in our case "XML article", place the respective XML document into the body, and embed JavaScript- and CSS-files. This then creates a Drupal node which transforms the document into XHTML.

4. OUTLOOK

The current status of full operability was achieved in about half a year by a team of six researchers/designers/programmers. Of course the main contribution is the scholarly TEI-encoding by the editor of the electronic edition which requires a more sustained scientific effort.

Future additions will include user annotations and images of pages tied to the already existing page links. This poses no great technical challenges. Somewhat more complex future developments include the addition of advanced search and indexing functionalities to the edition.

More demanding next steps planned are (1) an alternate stylesheet for transformation into the EPUB-format, which is a free and open e-book standard defined by the International Digital Publishing Forum (IDPF) [7], and (2) the production of a generic "TEI-for-Drupal"-module. This module will have to include some customization options for feature visualization as well as take into account a more complete set of TEI tags, ideally the complete TEI-Lite set. It will have to address, for instance, automating the process of generating a JavaScript array which maps print pages to in-file links, needed for across-chapter navigation. In the end, such a tool would allow humanities scholars with little experience in the more technical aspects of XSL and JavaScript programming to publish their TEI-encoded text editions in a standards-compliant, lightweight, sustainable and accessible framework.

5. ACKNOWLEDGMENTS

We gladly acknowledge the work of Dmitrij Funkner, Mohammed Abuhaish, and Roman Kominek who took on the implementation and documentation of this interdisciplinary project.

6. REFERENCES

[1] Bernard Barbiche. *L'édit de Nantes et ses antécédents (1562–1598)*, École nationale des chartes, 2003; 2e éd., 2009 (Éditions en ligne de l'École des Chartes, 5), dir. B. Barbiche, http://elec.enc.sorbonne.fr/editsdepacification.

[2] François-Joseph Bérardier de Bataut. *Essai sur le récit, ou Entretiens sur la manière de raconter*, Paris: Charles-Pierre Berton, 1776.

[3] Drupal project page, http://www.drupal.org.

[4] Penny Eley et al. (Eds.). *Partonopeus de Blois: An Electronic Edition*, HriOnline, Sheffield, 2005, http://www.hrionline.ac.uk/partonopeus/.

[5] Vincent van Gogh: *The Letters*, http://www.vangoghletters.org/vg/.

[6] Thierry Hoquet and Pietro Corsi. *Buffon et l'histoire naturelle*, édition en ligne, http://www.buffon.cnrs.fr/.

[7] International Digital Publishing Forum, EPUB Specifications, http://www.idpf.org/specs.htm.

[8] Christof Schöch. Édition électronique de Bérardier de Bataut, *Essai sur le récit* (1776), 2010, http://www.berardier.org.

[9] Susan Schreibman, Ray Siemens, John Unsworth (eds). *A Companion to Digital Humanities*, Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/.

[10] C. M. Sperberg-McQueen. *Textual Criticism and the Text Encoding Initiative*. December 1994, MLA '94, San Diego. http://xml.coverpages.org/sperb-mla94.html.

[11] Text Encoding Inititative, http://www.tei-c.org/Guidelines/.

[12] Text Encoding Initiative, List of projects, http://www.tei-c.org/Activities/Projects.

[13] W3C. XSL Transformations (XSLT) Version 2.0, W3C Recomm. 23 January 2007, http://www.w3.org/TR/xslt20.