Specifying a TEI-XML Based Format for Aligning Text to Image at Character Level

Alexei Lavrentiev, Dominique Stutzmann, Yann Leydier

Specifying a TEI-XML Based Format for Aligning Text to Image at Character Level

Alexei Lavrentiev, ICAR Research Lab, CNRS
<alexei.lavrentev@ens-lyon.fr>

Yann Leydier, Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France <yann@leydier.info>

Dominique Stutzmann, Institut de Recherche et d'Histoire des Textes, CNRS <dominique.stutzmann@irht.cnrs.fr>

Abstract

This papers presents an experience of specifying and implementing an XML format for text to image alignment at word and character level within the TEI framework. The format in question is a supplementary markup layer applied to heterogeneous transcriptions of medieval Latin and French manuscripts encoded using different "flavors" of the TEI (normalized for critical editions, diplomatic or palaeographic transcriptions). One of the problems that had to be solved was identifying "non-alignable" spans in various kinds of transcriptions. Originally designed in the framework of a research project on the ontology of letter-forms in medieval Latin and vernacular (mostly French) manuscripts and inscriptions, this format can be of use for all kinds of projects that involve fine-grain alignment of transcriptions with zones on digital images.

Table of Contents

Project Background[1]

The problems of fine-grain text to image alignment and its recording in standard interchangeable XML format were brought about by the Oriflamms Research Project [http://oriflamms.hypotheses.org/

[1]The results presented in this paper were obtained in the framework of the Oriflamms Research Project (ANR-12-CORP-0010-02) financed by the French National Research Agency (ANR). The authors are also grateful to the ASLAN project (ANR-10-LABX-0081) of Université de Lyon, for its financial support within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) of the French government operated by the ANR.

] aiming at establishing an ontology of forms and at analyzing the graphical structures of Medieval scripts (Latin and vernacular). Such an ontology would be useful for further studies in linguistics and various disciplines concerned with the history of scripts (palaeography, epigraphy, diplomatics, etc.). It would also provide valuable data for the development of software for optical recognition of medieval handwriting (HTR). One of the main project deliverables is a corpus of transcriptions aligned to facsimile image zones at word and character level, encoded in a standard interchangeable format (the choice was made to use TEI XML) and distributed under a free license[2].

To attain its goal, the project brought together researchers and engineers from several research labs in the fields of Humanities (palaeography, epigraphy and linguistics) and computer science (image analysis), as well as from a commercial enterprise specialized in optical character recognition. The partners of the project contributed a number of manuscript or inscription transcriptions and corresponding digital images of the source documents. All the transcriptions were TEI XML encoded but the precise markup schemas varied considerably depending on their purposes (historical, linguistic or literary studies), on the philological traditions followed ("diplomatic" vs. "critical" text editing) and on their age (from late 1990-s to 2013).

Input Transcription Types and Formats

Before presenting technical issues that had to be solved in the project, we will briefly describe the basic transcription types used in paper editions, and the ways they can be marked up and combined using computer technologies. The details and examples provided are mostly taken from the editions of Medieval Latin and French manuscripts but similar traditions exist for other European languages.

Palaeographic (Allographetic) Transcription

Transcriptions of this kind are also called "imitative" but we prefer the term of "allographetic" that corresponds to the linguistic status of letter variants in the framework of the theory of graphemics [Coulmas 1999]. [3] This kind of transcription represents very faithfully the aspect of the source document writing. It distinguishes major letter variants, like "long" and "round" *s* (ſ *vs.* s) or "round" and "straight" *r* (ꝛ *vs.* r), reproduces abbreviation markers, original punctuation and word segmentation. As a rule, allographetic transcriptions reproduce (or represent somehow) the layout of the original document (including line breaks, scribal correction markers, marginalia, etc.).

Allographetic transcriptions are rare in paper editions and can mostly be found in handbooks and albums for palaeographers (e.g. [Koschwitz 1879] and [Careri et al. 2001]). In electronic textual editing "pure" such transcriptions are also very rare: the original manuscript transcriptions of the Charrette Project [http://www.princeton.edu/~lancelot/ss/] are one the few examples that we could find. More often, some allographetic features can be found in multi-layer electronic transcriptions that will be discussed later.

This kind of transcription is the easiest to align with image zones. However, problems may arise in the process of word tokenization, as the white spaces that appear in manuscripts (and reproduced as they are in allographetic transcriptions) do not always correspond to word limits in terms of modern linguistic analysis. Another problem may concern abbreviation markers, some of which (but not all) are superscribed over the baseline characters. This has however more to do with the technique of image segmentation than with the transcription itself.

Diplomatic Transcriptions

Diplomatics is a scholarly discipline that studies the tradition, the form, and the production of written acts (especially historical documents) [VID 1997: 21]. In diplomatic editing, it is important to preserve as much information as possible from the source document without overcharging it

[2] *I.e.* conform to the Free Software Foundation licensing principles (http://www.fsf.org/licensing).
[3] In disciplines like Epigraphy this kind of transcription is usually referred to "diplomatic" (*vs.* "interpretative"), however, we prefer to use this term for the intermediate transcription type described in the next section.

with information not relevant for its interpretation (such as, typically, letter variants, or allographs). The practice of diplomatic editions has been far from being stable along its history and different national traditions. Some normalization efforts were undertaken by the *Commission internationale de diplomatique* (http://cidipl.org) but their recommendations [Bautier ed. 1984] are very general and allow considerable variation in local practices. It is nevertheless possible to point out some typical features of diplomatic transcriptions. For instance, the abbreviations are expanded but the letters supplied are typographically marked (using italics or some kind of braces). The original document layout is usually preserved. The punctuation and word segmentation may be normalized according to modern typographical rules. Some diplomatic editions also use character disambiguation (introducing *u / v* and *i / j* phonetic distinction and adding some diacritics that help reading the text) and capitalize the proper names. In some cases the editor may supply a missing word or part of word, or point out to a repeated word or phrase but these emendations are always clearly marked typographically. As an example of modern diplomatic edition we can quote *Les plus anciens documents linguistiques de la France* published online by Zurich University (http://www.rose.uzh.ch/docling).

Diplomatic transcriptions can in principle be aligned to image zones, provided that the abbreviation expansions, scribal corrections and editorial emendations are properly tagged. The punctuation marks may need to be ignored if the punctuation is modernized in the transcription. The neutralization of allographs, disambiguation of some characters and capitalization of proper names do not affect the alignment but need to be addressed in the markup model.

Normalized (Critical) Transcriptions

Transcriptions of this kind are the most widespread in the field of literary text editing. As far as the Old French texts are concerned, the first “normalization” recommendations (although the term was not used) were published by a commission of *Société des Anciens Textes Français* in 1926 [Roques 1926]. These recommendations were precise enough at some points (like using diacritics) but a number of questions (like word separation) were not mentioned at all. In early 2000s, *École Nationale des Chartes* published a series of recommendations for editing various kinds of medieval texts (Latin and vernacular, literary and documentary) [Vielliard ed. 2001, Guyotjeannin ed. 2001 and Bourgain ed. 2002] but they are not always followed in practice.

Without speaking of “reconstructive” (or “Lachmannian”[4]) critical editions based on the comparison of multiple witnesses, which are by definition not alignable to an image of a single manuscript, even relatively “faithful” normalized transcriptions are hardly usable for the purposes of alignment. Whereas major editorial emendations to the source document are usually clearly marked (e.g. supplied words or letters are placed into square brackets), some information important for alignment is lost: the expansions of abbreviations are unmarked, as are the scribal corrections, and the manuscript page layout is rarely preserved (except, to a certain extent, the verse lines). However, thanks to digital markup, critical transcriptions may become more “alignment-friendly”. In this case, they should not be referred to as “plain critical” but rather as “hybrid” or “multi-layer” transcriptions.

Hybrid Digital Transcriptions

The possibility offered by computer technologies to have both abbreviation markers and expansions, corrected or regularized and original erroneous or irregular forms is one of the main arguments in favor of producing digital editions. The TEI provided special elements for these purposes since its very early versions. Initially, the alternative versions were encoded by means of attributes (e.g. the `<expan>`[5] element had an “abbr” attribute), but the P5 version introduced a new `<choice>` element in order to group alternative encodings of the same segment of a text. This mechanism is much more flexible and powerful than the old attribute-based system but it may require more complex processing. TEI allows using more than two child elements inside `<choice>`, and `<choice>`s may nest. Some restrictions had to be added to the content model of `<choice>` for the alignment project, such as allowing only two child elements. Some more restrictions and requirements for the alignable transcription format will be described in the section called “Target Format”.

[4] Named after Karl Lachmann (1793-1851), a German classical philologist considered to be the founder of the method.
[5] Unless otherwise stated, all XML elements cited in this article are those defined by the TEI (http://www.tei-c.org/ns/1.0 namespace).

The actual transcriptions used in the project can all be qualified as “hybrid” but precise encoding practices differed considerably. Characteristic features of three of them are listed below.

Charrette Transcriptions

The Princeton Charrette Project manuscript transcriptions were initially encoded in TEI P3 SGML with extensive use of SGML entities for “non-ASCII characters”. These transcriptions were purely allographetic. They were automatically converted to XML (extended TEI P4X) in 2002, and are currently available in this format under a Creative Commons BY-NC-SA 2.5 License [http://creativecommons.org/licenses/by-nc-sa/2.5/] from the project legacy website [http://www.princeton.edu/~lancelot/ss/]. These transcriptions provide information on allographs grouping (e.g. that “round” and “long” *s* are variants of the same letter), and supply expansions for some abbreviations, but the latter have never been checked and contain many errors. The following example illustrates a verse line transcription from the Charrette project:

```
<l n='1' id='MS-A-196-r-a-1' key='FU-31'>
  <chr_large size="8" color="" detail="">A</chr_large> un jor
  dunea<chr_var letter="s" var="long">&#x222B;</chr_var>cen<chr_var
    letter="s" var="long">&#x222B;</chr_var>i<chr_abbr type="" class=""
    expan="" cert="no">[apost]</chr_abbr>o
</l>
```

One can see that project specific elements are used for abbreviations, allographs, and large capitals, and that no expansion is provided for the abbreviation in the end of the line.

Graal Multi-Layer Transcriptions

The digital edition of the *Queste del Saint Graal* (manuscript Lyon, Bibliothèque Municipale, P.A. 77) [Marchello & Lavrentiev ed. 2013] allows downloading its TEI XML source files under a Creative Commons BY-NC-SA 3.0 License [http://creativecommons.org/licenses/by-nc-sa/3.0/]. The Old French part of the edition is encoded in TEI P5 XML using the BFM-MSS extension. This extension includes, in turn, three extension elements defined by the Medieval Nordic Text Archive Initiative [http://www.menota.org] Project (MENOTA)[6]:`<me:norm>`, `<me:dipl` and `me:facs` that are wrapped into `<choice>` and placed inside `<w>` that tag every word or punctuation mark[7] of the text, as shown in the example below:

```
<w type="NOMpro" xml:id="w106_000286">
  <choice>
    <me:norm>Lancelot</me:norm>
    <me:dipl>lanc<ex>elot</ex></me:dipl>
    <me:facs>lan<bfm:mdvAbbr>c&#x0305;</bfm:mdvAbbr></me:facs>
  </choice>
</w>
```

The three Menota elements correspond roughly to the three transcription types described above. One can observe that the initial *l* is capitalized in the normalized layer (because *Lancelot* is a personal name), that the abbreviation marker in the form of horizontal bar placed above the *c* letter is represented by the corresponding Unicode character in the allographetic layer, and that the expansion letters are tagged with `<ex>` in the diplomatic layer.

The format of the *Graal* edition has the advantage of being explicit and easy to process, but it is very verbose and is not “TEI conformant”, as defined in in the TEI Guidelines [[TEIP5]: section 23.4]. [8] It should be noted that the “facs” (*i.e.* allographetic) transcription layer is only provided for the first nine columns of the text, so different alignment rules had to be applied to the rest of the text.

[6]The BFM-MSS proper extension elements are not relevant for the scope of this article and will not be described here. Complete ODD documentation of the BFM-MSS encoding schema is available at http://portal.textometrie.org/bfm/files/BFM-MSS-ODD.xml
[7]The `<pc>` element was not available at the time when the encoding schema of the project was defined.
[8]The Graal transcriptions qualify as a TEI extension and are close to being TEI conformable.

Fontenay Charters Transcriptions

The Fontenay charters were encoded in TEI P5 XML on the basis of a normalized edition, then enhanced with allographetic features, focussing on abbreviations and the distinctive allographs which were already identified as relevant for the research [Stutzmann 2011: 253-5]. Its origins make this transcription uneven: tagged named entities and normalized punctuation and capitalization, but with line breaks, abbreviations, and allographs. A fragment of the input transcription is provided below:

```
<lb n="14"/> Actum est <app>
  <lem>hoc</lem>
  <rdg wit="B" rend="omis"/>
</app> <placeName>Edu#</placeName> in plena #ynodo <date>
  <app>
    <lem>.XV.</lem>
    <rdg wit="B">XV°</rdg>
  </app> kl. julii</date>, laudanti<choice>
  <expan>bus</expan>  <abbr>b;</abbr>
</choice> archidiaconi<choice>
  <expan>bus</expan>
  <abbr>b;</abbr>
</choice>
<persName> Gaufrido</persName>,
```

One can note simultaneous use of apparatus elements, name and date tags, abbreviation marks and expansions, and the presence of modernized punctuation (which is not tagged). The use of whitespaces in elements with mixed content is relevant for tokenization.

Target Format

The scope of this format is to enable text to image alignment at word and character level avoiding as much as possible any manual editing of the input transcriptions. It defines a set of layout elements necessary for positioning a text fragment on the page surface, a number of rules for distinguishing alignable and non-alignable elements of the transcription, and the mechanism for recording the alignment at word and character level.

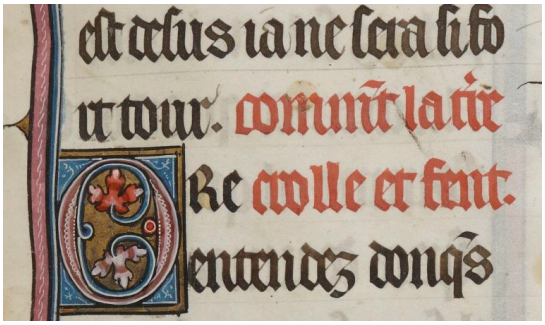
Layout Markup and Linearity Conflicts

In order to associate a transcription of a page with an image file, a <pb/> must be provided at the beginning of each transcribed page. It must have an `xml:id` attribute for stand-off alignment and may have a `facs` attribute for direct linking to the corresponding image file or to the corresponding `zone` element in the `facsimile` section.

A `<cb/>` must be used in the beginning of each text column if there are more than one of them on a page.

A `<lb/>` must be used in the beginning of each text line. It must have an `n` attribute indicating the line number. In most cases this attribute may be generated automatically but sometimes the physical organization of text segments in lines does not correspond to the logical order of the text structure. This actually happens at the border of some text divisions where the title of the new division (often written with red ink, as a *rubric*) overlaps with the end of the previous division (See Figure 1, “Rubric Overlapping with Text Divisions”).

Figure 1. Rubric Overlapping with Text Divisions



Example of rubric overlapping with a frontier of text divisions. Manuscript Paris, BnF, fr. 574, fol. 83r. Image obtained from the Gallica Virtual Library (http://gallica.bnf.fr [http://gallica.bnf.fr/ark:/12148/btv1b84526412/f179.item])

In this case, the linearity of the transcription follows the logical order of the text, and two `<lb/>` elements with the same value of the `n` attribute are used in the rubric and in the text body, as in the example below:

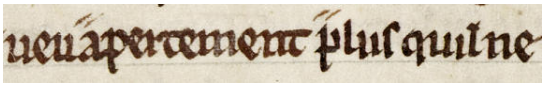
```
<div>
   <p>[...]
    <lb n="4"/> est desus ia ne sera si fo<lb break="no" n="5"/>rt
    tour .</p>
</div>
<div>
  <head>comm<ex>en</ex>t la t<ex>er</ex>re
    <lb type="rubrique" rend="align(right)" n="6"/> crolle et fent.</head>
    <p>
      <lb n="6"/> <hi rend="lettrine">O</hi>re
      <lb n="7"/> entendez donq<ex>ue</ex>s [...]
    </p>
   </div>
```

Thanks to this simple mechanism, the alignment software can find the physical lines corresponding to the transcription of rubrics and similar cases. The `type` attribute on the first occurrence of hte “duplicated” `<lb/>` allows sorting out the line breaks that different from those in the “main” text flow. The optional `rend` attribute may be used to indicate the position of the following text in the physical line. In our example, the text “crolle et fent.” is placed to the right of “Ore” and thus overlaps with the logical order of the transcription.

Another case where the linearity of the transcription does not correspond to the physical order of lines in the manuscript is that of interlinear or marginal additions. These must be marked up using the `<add>` tag and are considered non-alignable at the current stage of the project, as the image analysis software only detects the regular lines of the text.

A yet more complex situation is created in the case of scribal corrections of word order usually indicated by special transposition markers (See Figure 2, “Transposition”).

Figure 2. Transposition



Example of transposition. Manuscript Lyon, B.M., P.A. 77, col. 163d. Image provided by Bibliothque Municipale de Lyon. The corrected text reads “ueu plus apertement qu'il ne”

If the transcription follows the “final” text order (which is usually the case), special markup is necessary in order to ensure the possibility of alignment. A combination of `<ptr>` and `<seg>` elements is proposed for this purpose in the Oriflamms project:

```
<ptr type="transposition-orig" target="#tps106_163d2_1"/>
<w type="ADVgen" xml:id="w106_006642">plus</w>
<seg type="transposition-target" xml:id="tps106_163d2_1">
  <w type="ADVgen" xml:id="w106_006643">apertement</w>
</seg>
```

The following XSL templates can be used to rearrange the text in the “physical” order:

```
<xsl:template match="tei:ptr[@type='transposition-orig']">
  <xsl:apply-templates select="id(substring-after(@target,'#'))"/>
</xsl:template>

<xsl:template match="tei:seg[@type='transposition-target']"/>
```

Manuscript pages can contain inscriptions that do not belong to the text being transcribed. These include catchwords and posterior annotations (such as library shelf marks). In some cases these elements may be transcribed (using `<fw>` and `<add>` tags) but in most cases they are not, and the alignment software has to guess which lines on the manuscript page are “suspect” and dropped from further analysis.

Alignable and Non-Alignable Elements

In addition to the transription *stricto sensu*, XML documents may contain a whole number of text containing elements that do not correspond to any graphical object in the source manuscript. These include the metadata recorded in the `<teiHeader>`, editorial notes (`<note>`), text supplied by the editor for missing or illegible passages (`<supplied>`). These are of course non-alignable. On the contrary, the elements `<del>` and `<surplus>` contain parts of text which should not appear in the edition (e.g. repeated or expunctuated words), but should be taken into account by the alignement. The transcription of these text segments had to be added to the source documents where appropriate.

Some transcriptions include critical apparatus providing variants from different manuscripts of the same text tradition (*cf.* the Fontenay transcription fragment provided above). These are encoded using `<app>` with `<lem>` and `<rdg>` child elements. In this case, the alignable text content is situated in the `<lem>` or `<rdg>` where the `wit` attribute contains a reference to the base manuscript of the transcription.

Whereas `<choice>` elements are used, only one of their child elements is alignable: `<sic>` (and not its `<corr>` sibling), `orig` (and not its `<reg>` sibling), `<abbr>` (and not its `<expan>` sibling). However, if `<reg>` and `<expan>` are used alone, they are alignable at word level, even though some characters in the transcription will not match those on the image. The `<corr>` element used alone may be alignable at word level but as is contains a segment of text somehow corrected by the editor, there is no guarantee that the characters contained by this element have any correspondence on the image.

The abbreviations may be particularly hard to align, as some of the markers occupy a position on the horizontal text line (as baseline or spacing superscript characters) and others are placed above other characters (as combining diacritics or superscript letters). The former are alignable, the latter are not. When the transcription represents the abbreviation markers directly, the Unicode areas labelled as “combining” can be used to identify the non alignable characters. When the transcription only contains an expansion with the supplied letters marked up using the `<ex>` tag, the only way to identify the alignable expansions (which are a minority) is to provide a list of them. Here is the list that has been used in the project so far:

- “9”, or “overturned *c*” on the baseline for “con”, “com”, or “cum” (ꝯ),
- “7”, or “barred 7” for “et” (⁊ or  (MUFI)),

- “÷” for “est” (∻),
- angular tilde for “er” (͛ ou  (MUFI)),
- double curb tilde for “ur” (᷑ ou  (MUFI)),
- “9-shaped” tilde for “us” (ꝰ).

Word and Character Level Tokenization

The `<w>` and `<pc>` tags are used on every alignable word and punctuation mark. All of them are equipped with an `xml:id` attribute which is used for alignment. The same technique is applied at character level using the `<c>` tag. The tokenization at both levels is performed automatically using pre-processing XSLT stylesheet librrary.

Image Markup and Linking

The results of the alignment software are recorded in separate files using standard TEI digital facsimile encoding mechanism with `<zone>` elements equipped with attributes indicating coordinates and with `xml:id`.

The linking is between the transcriptions and image zones is recorded in separate TEI files containing `<link>` and `<linkGrp>` elements joining references to the identifiers of words or characters in the transcription and of zones in the facsimile.

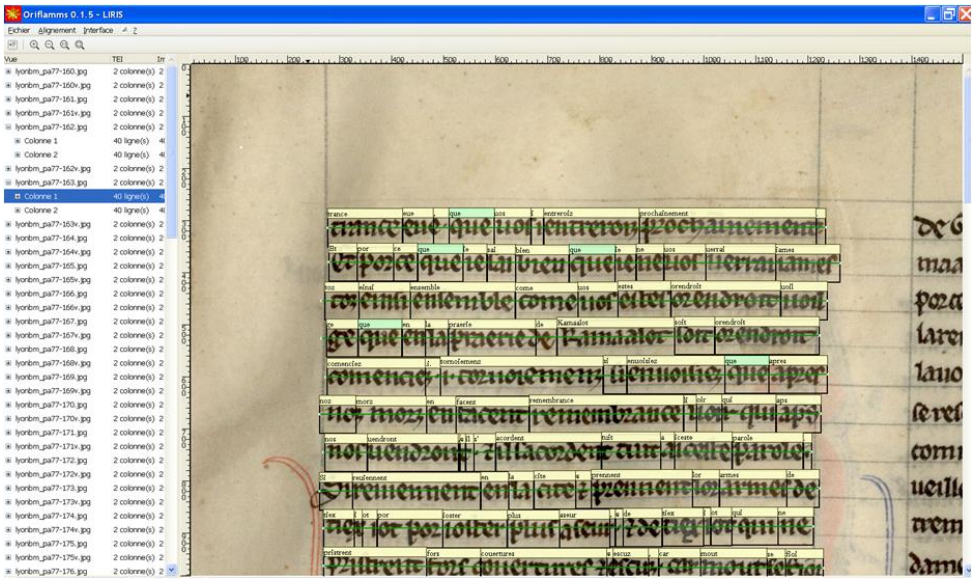
Processing and Implementation: Oriflamms Alignment Software

The Oriflamms software allows the automated alignment of TEI words with images, the manual and automated validation and correction of the alignment [Stutzmann 2013 and Leydier 2014]. It also offers the possibility to display all image occurrences of a query string and to export comprehensive statistics on the validation and correction actions performed by the user.

To create an alignment project, one needs a TEI file and the set of images that are referred to in the `<pb>` elements. The software then automatically detects the bounds of text columns in the images using the count of `<cb>` elements. Median lines are then computed in the columns. If there are more lines on the image than there are `<lb>` elements, then the most uncommon lines are ignored in priority (e.g. catchwords or library shelf marks inscribed on the page). Words are not segmented since the spaces between them are at best irregular and often nonexistent. Along each median line a signature is computed. It describes the coarse shape of the strokes: dots, curves facing left or right, vertical lines above, bellow or crossing the median line. The same kind of signature is extracted from Unicode text lines using a lookup table. The signatures of Unicode and image text lines are compared with the Levenshtein distance from which the alignment is automatically inferred. Because of the kerning, vertical segments are unsuited to separate words and characters, so we use curves that follow the lightest path (considering that the text is darker than the background).

The graphical user interface displays the alignment on the images with boxes on top of which sits the transcriptions. A mouse click on a words makes it cycle through three states of validation: correct, erroneous and unvalidated.

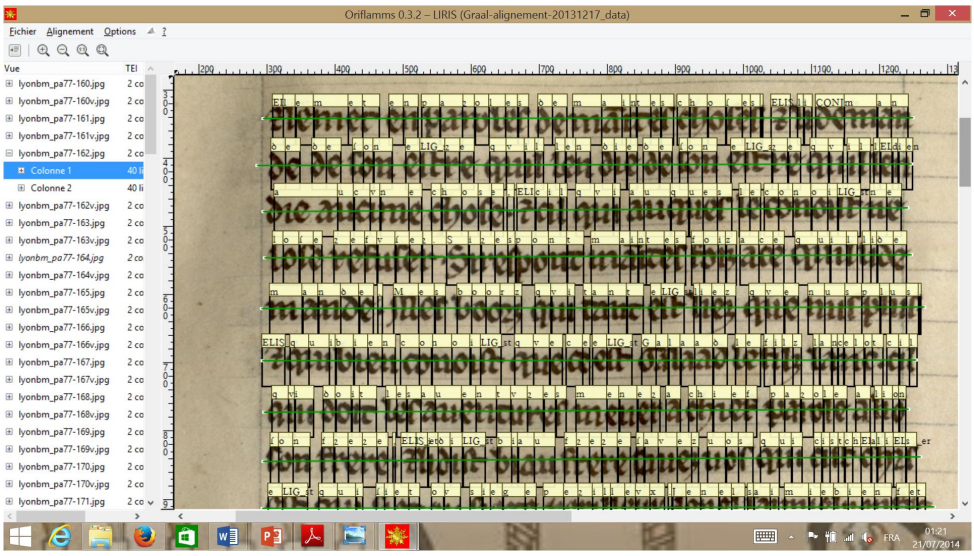
Figure 3. Oriflamms Software Word Alignment Validation and Correction Interface



This interface allows verifying and correcting word level alignement on the manuscript page image.

The same interface allows validating and correcting text-to-image alignment at character level:

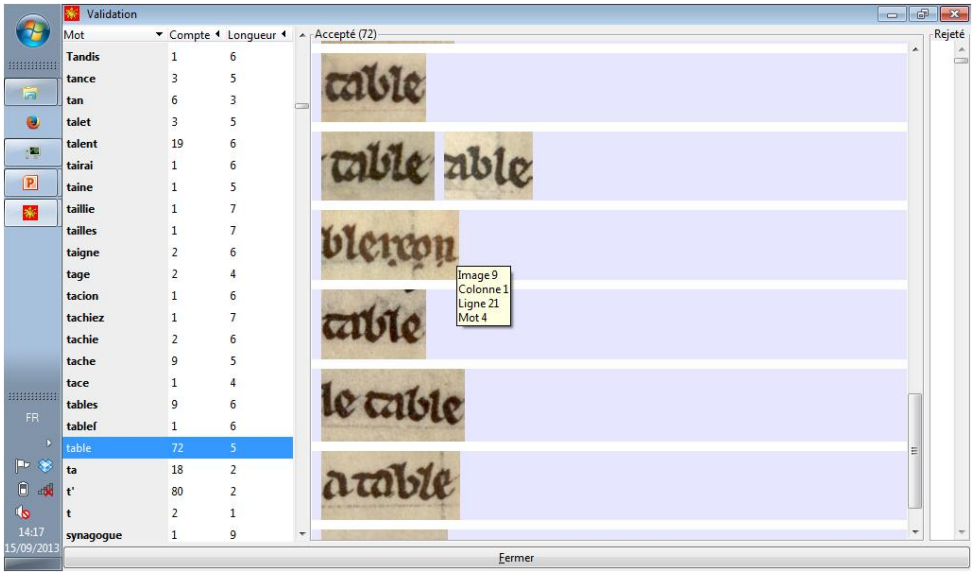
Figure 4. Oriflamms Software Character Alignment Validation and Correction Interface



This interface allows verifying and correcting character level alignement on the manuscript page image.

A second way of validating is also offered *via* a tabular view. It displays a list of all the words that can be sorted alphabetically, by the number of occurrences or by length.

Figure 5. Oriflamms Software Tabular Word Alignment Validation Interface



This interface allows validating word level alignement by word lists.

When a word is selected, all its image occurrences are displayed in several clusters sorted by graphical similarity. The user may click (or click and drag) to tag the wrongly aligned occurrences. The manual validation of the alignment can be expanded automatically with simple rules such as: “a word lying between two validated words is valid”.

On the image view, the user can enable the edit mode that allows to correct the median lines and the word alignment. In edit mode, the words' bounding boxes are rectangular to allow a simple correction of the frontiers. The curvy frontiers are recomputed automatically when a word's alignment is modified. When a frontier is edited, all preceding or following unvalidated or wrongly aligned words are realigned consequently.

Statistics can be exported to a spreadsheet. They describe the alignment correctness of validated word including rate for each letter beginning or ending words. The average correction (in pixels) of the front and back frontiers is also displayed. These statistics allow us to analyse which words, letters or sequences of letters are baldy aligned in order to improve our algorithms.

Conclusion

The research project that inspired the reflection on the problems of text to image alignment is still going on, and the proposed markup solutions presented above are to a certain extent “work in progress”. We hope that sharing this experience at the Balisage conference may result in the improvement and consolidation of the proposed format and of the processing software that will eventually benefit the whole community of scholars working with images and transcriptions of text-bearing objects.

Bibliography

[Bautier ed. 1984] Bautier, Robert-Henri, ed. *Folia Caesaraugustana, vol. 1 : Diplomatica et Sigillographica. Travaux préliminaires de la Commission internationale de diplomatique et de la Commission internationale de sigillographie pour une normalisation des éditions internationales des éditions de documents et un Vocabulaire internationale de la diplomatique et de la sigillographie.* Zaragoza. Institución “Fernando el Católico”. 1984.

[Bourgain ed. 2002] Bourgain, Pascale and Vielliard, Françoise (ed.) *Conseils pour l'édition des textes médiévaux. Fascicule III, Textes littéraires*. Paris. CTHS, École nationale des chartes. 2002.

[Careri et al. 2001] Careri, Maria, Fery-Hue, Françoise, Gasparri, Françoise, Hasenohr, Geneviève, Labory, Gillette, Lefèvre, Sylvie, Leurquin, Anne-Françoise, and Ruby, Christine. *Album de manuscrits français du XIIIe siècle. Mise en page et mise en texte.* Rome. Viella. 2001

[Coulmas 1999] Coulmas, Florian. *The Blackwell Encyclopedia of Writing Systems*. Oxford, OX, UK ; Cambridge, Mass., USA: Blackwell Publishers. 1999. doi: 10.1111/b.9780631214816.1999.x.

[Guyotjeannin ed. 2001] Guyotjeannin, Olivier, ed. *Conseils pour l'édition des textes médiévaux, Fascicule II, Actes et documents d'archives*. Paris. CTHS, École nationale des chartes. 2001

[Koschwitz 1879] Koschwitz, Eduard. Les plus anciens monuments de la langue française publiés pour les cours universitaires par Eduard Koschwitz. Henninger. Heilbronn. 1879. http://gallica.bnf.fr/ark:/12148/bpt6k3733294.

[Leydier 2014] Leydier, Yann, Eglin, Véronique, Bres, Stéphane, and Stutzmann, Dominique. "Learning-Free Text-Image Alignment for Medieval Manuscripts", *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014, p. 363#368. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6981046, 10.1109/ICFHR.2014.67

[Marchello & Lavrentiev ed. 2013] Marchello-Nizia, Christiane and Lavrentiev, Alexei, ed. *Queste del saint Graal. Édition numérique interactive.*[online]. 2013. http://catalog.bfm-corpus.org/qgraal_cm

[Roques 1926] Roques, Mario. "Établissement des règles pratiques pour l'édition des anciens textes français et provençaux. Rapport de la 2e commission." *Romania*, vol. 52, 1926. http://gallica.bnf.fr/ark:/12148/bpt6k16060g

[Stutzmann 2011] Stutzmann, Dominique. "Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin#?", *Codicology and Palaeography in the Digital Age 2*, Norderstedt. 2011, p. 247-277 http://kups.ub.uni-koeln.de/4353/.

[Stutzmann 2013] Stutzmann, Dominique. "Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS", *Document numérique*, 16/3, 2013, p. 81-95.

[TEIP5] The TEI Consortium, *TEI P5: Electronic Text Encoding and Interchange*. 2015. http://www.tei-c.org/Guidelines/P5/

[VID 1997] *Vocabulaire international de la diplomatique*. València. Universitat de València. 1997. http://www.cei.lmu.de/VID/

[Vielliard ed. 2001] Vielliard, Françoise and Guyotjeannin, Olivier, ed. *Conseils pour l'édition des textes médiévaux, Fascicule I, Conseils généraux.* Paris. CTHS. École nationale des chartes. 2001.