Automatically Encoding Encyclopedic-like Resources in TEI

Mohamed Khemakhem, Laurent Romary, Simon Gabay, Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo

Automatically Encoding Encyclopedic-like Resources in TEI

Mohamed Khemakhem[1,2,3], Laurent Romary[1,2,4]
{name.surname@inria.fr}
[1] Inria ALMAnaCH, Paris
[2] Centre Marc Bloch, Berlin
[3] Paris Diderot University, Paris
[4] BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

Simon Gabay
{surname.name@unine.ch}
Université de Neuchâtel - Institut de littérature française

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo
{name.surname@univ-montp3.fr}
Univ Paul Valéry Montpellier 3, CNRS, PRAXILING UMR 5267

Encyclopedic content exists in different forms of paper-based resources and remains to a very large extent not exploited given the limited alternatives to extract information from the corresponding digitized material. Where TEI has provided mechanisms for encoding such a content, recent works have yielded advanced techniques to apply these mechanisms automatically on raw resources, however clear recommendations for the encoding of such content is still lacking.

As current digitization projects concern encyclopedias, other projects are undertaken to digitize textual resources with heritage significance and with similar features in terms of layout and indexed access. In this paper we introduce new lexicographic models, customizing the TEI schema, we extend existing ones and we present their application on two categories of resources, currently under development to be made available for public use:

- Encyclopedic sections of early, out of copyright, versions of a reference French Dictionary (Petit Larousse Illustré),
- Manuscripts auction catalogues, which are old resources for referencing and semantically describing manuscripts for sale.

The definition of these models is based on the significant similarities noticed at different levels of the described information in such entry-based resources. We have employed elements from the TEI dictionaries module to enable the segmentation of morphological and semantic information encompassed by respectively <form> and <sense> elements, already extracted by dedicated existing models in GROBID-Dictionaries[1]'s architecture. In fact, each entry in both types of resources carries two main encyclopedic information. First, the name or the label of a concept, coupled with one or more of its extended form and coming sometimes with a brief description. Then, a second block comes to establish an exhaustive description of the concept and to

[1] https://github.com/MedKhem/grobid-dictionaries/

present related information such as bibliography or pricing. In addition, we have used a customized version of the <entry> element to allow the generic encoding of entry numbering (see second excerpt in the table below).

TEI encoding	Excerpt from Dictionary Encyclopedic Section in Petit Larousse Illustré (1948)
<entry> <form type="lemma"> <persName>ABERDEEN</persName> <addName>(G. H. Gordon, comte d')</addName><pc>,</pc> <desc>homme d'Etat anglais, né à Edimbourg</desc> </form> <pc>.</pc> <sense> <def>Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860)</def> </sense> <pc>.</pc> </entry>	**ABERDEEN** [*abɐrdin'*], v. d'Ecosse, ch.-l. de comté; port sur la mer du Nord; 170.000 h. Université. **ABERDEEN** (G. H. Gordon, *comte d'*), homme d'Etat anglais, né à Edimbourg. Premier ministre en 1852, il conclut avec la France une alliance contre la Russie (1784-1860). **ABER-VRACH**, fl. côtier du Finistère (Atlantique); 34 kil. Station marémotrice d'essai. **ABGAR**, nom de huit rois d'Edesse, en Mésopotamie (132 av. J.-C.-216 apr.). **ABIA**, roi de Juda, fils de Roboam, vainqueur de Jéroboam, roi d'Israël (957-955 av. J.-C.). **ABIDJAN**, ch.-l. de la Côte-d'Ivoire (A.-O. F.), sur une vaste lagune navigable; 15.000 h. **ABIMÉLECH** [*lèk*], fils de Gédéon. Il devint Juge d'Israël, après avoir fait égorger ses frères; il établit son pouvoir sur Sichem et fut tué au siège de Thèbes, en Palestine (vers 1100 av. J.-C.). **ABIRON**, lévite qui fut englouti dans la terre avec Coré et Dathan, tous trois révoltés contre Moïse et Aaron (*Bible*).

TEI encoding	Excerpt from a Manuscripts Auction Catalogue (1889)
<entry> <num>49</num> <form type="lemma"> <surName>Kourakin</surName> <addName>(le prince Alexis B.),</addName> <desc> frère du précédent, homme d'Etat russe.</desc> </form> <sense> <pc>-</pc> <def> <bibl>Billet auto sig., en francais, à M. Monférand, 1 p, in-8.</bibl> <num type="price">2 »</num> </def> </sense> </entry> <entry> <num>54</num> <form type="lemma"> <surname>Lassalle</surname> <addName>(A.-Ch.-L. de)</addName>, <desc>le plus brillant général de cavalerie des guerres de la République et de l'Emplre, né à Metz, tué à la bataille de Wagram</desc> </form> <sense> .- <def> <bibl>L. a. s. au général Dugua; 1 p. in-f.</bibl> <num type="price">10 »</num> </def> <note>Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite affaire, tué 10 mameloucks et 40 arabes, etc.</note> </sense> </entry>	49 **Kourakin** (le prince Alexis B.), frère du précédent, homme d'Etat russe. — Billet aut. sig., en français, à M. Monférand, 1 p. in-8. 2 » 50 **Labanoff** (le prince Alex.), célèbre général et écrivain russe, historien de Marie Stuart. — L. a. s., en français, 1835, 1 p. in-4. 3 » 51 **Ladislas IV**, roi de Pologne, célèbre par ses succès contre les Russes, époux de Marie de Gonzague. — L. sig., en latin, au cardinal de Montalte; Varsovie, 1645, 1 p. in-f. 8 » 52 **Lafayette**, illustre général. — L. a. sig. de ses initiales à M. Masclet; Washington, 13 août 1825, 1 p. 1/4 in-4. Un peu fatiguée. 15 » Très-curieuse lettre sur le voyage qu'il fit en Amérique, de 1824 à 1825. « C'est avec de bien tendres regrets que je quitterai cette terre américaine, le bon, grand et heureux peuple des Etats-Unis auquel je suis amalgamé depuis près d'un demi-siècle, et qui vient encore de me combler de ses bontés. J'y ai vu les miracles de l'indépendance, de la liberté, égalité et *self government;* le problème des institutions républicaines a été résolu ici sur une grande échelle et jamais expérience n'a si bien réussi. » Il comptait retourner comme il était venu, sur un paquebot-poste, mais le peuple et le gouvernement en ont disposé autrement. On a donné le nom de *Brandywine* à une superbe frégate qui est chargée de le ramener en France. 53 **La Roncière** (Emile-*Clément* de), fils du général, condamné pour tentative de viol. — L. a. s. aux officiers et élèves de l'école de Saumur; Paris, mai 1836, 3 p. pet. in-4. 10 » Très-curieuse lettre toute relative à son procès. 54 **Lassalle** (A.-Ch.-L. de), le plus brillant général de cavalerie des guerres de la République et de l'Empire, né à Metz, tué à la bataille de Wagram. — L. a. s. au général Dugua; 1 p. in-f. 10 » Superbe lettre sur la campagne d'Egypte. Il profite du départ du général Desaix pour lui donner des nouvelles. Desaix lui laisse le commandement de la colonne qui doit poursuivre Mourad-Bey, et qui se compose de 400 hommes de cavalerie, 4 pièces de canon et 160 dromadaires. Le général Boyer a, dans une petite affaire, tué 10 mameloucks et 40 arabes, etc.

We employed the adapted and newly defined models to automatically process and label the encyclopedic information in a cascading fashion, as introduced in Khemakhem et al. 2017 and Khemakhem et al. 2018, by relying on the same sequence labeling machine learning technique.

We will present in-depth our progress in implementing morphological and semantic models and show through the results at each structuring level how such information could be uniformly encoded and automatically extracted from these resources.

References:

1. Mohamed Khemakhem, Luca Foppiano, Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. electronic lexicography, eLex 2017, Sep 2017, Leiden, Netherlands
2. Mohamed Khemakhem, Axel Herold, Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. GLOBALEX workshop at LREC 2018, May 2018, Miyazaki, Japan.
3. Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem, Laurent Romary. Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. GLOBALEX 2018 - Globalex workshop at LREC 2018, May 2018, Miyazaki, Japan.