Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, Lucie Rondeau Du Noyer

To cite this version:
Mohamed Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, et al.. Information Extraction Workflow for Digitised Entry-based Documents. DARIAH Annual event 2020, May 2020, Zagreb, Croatia. hal-02508549

HAL Id: hal-02508549
https://hal.archives-ouvertes.fr/hal-02508549
Submitted on 14 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem$^{1,2,3,4}$, Simon Gabay$^{5,6}$, Béatrice Joyeux-Prunel$^6$, Laurent Romary$^{2,3,7}$, Léa Saint-Raymond$^8$, and Lucie Rondeau Du Noyer$^9$

$^1$LittArts - UMR 5316, Grenoble
$^2$Inria - ALMA(na)Ch - Automatic Language Modelling and ANAlysis Computational Humanities
$^3$CMB - Centre Marc Bloch
$^4$Université Paris Diderot, Paris
$^5$Université de Neuchâtel
$^6$Université de Genève
$^7$BBAW - Berlin-Brandenburg Academy of Sciences and Humanities
$^8$Ecole normale supérieure, Paris, DMA

January 2020

1 Program

It is a 4 hours workshop:

- 30 mn: introduction of the workflow and the two tools
- 30 mn: setting up the tools on the machines
- 2 h: Two Parallel groups. Goal: Training a model for a dictionary and a model for a catalogue
- 30 mn: towards a digitisation pipeline
- 30 mn: Q&A

2 Instructors

- Mohamed Khemakhem
3 Target Audience

Scholars and humanists with basic notions of command lines and familiar with the TEI standard.

4 Abstract

The massive retro-digitisation of legacy paper resources in the last decade, along with the constant growth of compiled unstructured digital text material, have created an unbalanced situation where the existent ad hoc techniques for exploiting such resources are unable to cover the important stream of emerging corpora. In this workshop we address this issue and present an exploratory workflow implemented in two state of the art infrastructures for Information Extraction (IE) from documents with entry-based structure and diverse content.

IE in Digital Humanities (DH) has always been a serious challenge for researchers dealing with modern or legacy text resources [1, 2]. GROBID-Dictionaries\textsuperscript{1} is a project which has been launched to fill in this gap by accelerating the modelling and structuring of resources within the lexicography field. The first version of the machine learning infrastructure has been focused on structuring digitised dictionaries into TEI-compliant resources [3].

In GROBID-Dictionaries, the activation of cascading IE models follows an exploratory process based on the MATTER workflow [4]. Throughout a multi-stage annotation and curation process, a user of the tool discovers gradually the structure and the variation of the information in a target document.

\textsuperscript{1}\url{https://github.com/MedKhem/grobid-dictionaries}
In addition, and despite being initially designed for structuring dictionaries, the implemented approach and the tool’s pluggable models have shown enough flexibility to be applicable on a wider range of modern and legacy entry-based documents such as dictionaries [3, 5] (fig. 4), legacy address directories [6] (fig. 6), and large bibliographic collections [7].

The rising interest around the applicability of such an IE approach to certain entry-based documents [8, 9] and its fitness potential in larger document processing pipelines, triggered the creation of a second tool adapted for analysing and structuring a family of prints: Catalogues (fig. 5 & 7). Grobid-Cat\(^2\) is a fork of Grobid-Dictionaries that remains faithful to the exploratory cascading approach with more optimised combination of parsing models (fig. 2) and more adapted TEI encoding (fig. 3).

\(^2\)https://github.com/MedKhem/grobid-cat

Figure 2: RDA, No 67 (March 1881), lot No 55.
Figure 3: *RDA*, Output from Figure 1 processed with GROBID-Cat.

The proposed workshop represents an illustration of scholarly primitives practices employing a new generation of IE tools. Our goal is to familiarise interdisciplinary DH users with the MATTER workflow through two parallel hands-on sessions, each one dealing with a different category of digitised documents: dictionaries and catalogues. We rely on enhanced usability features [5] to ease the setup and the manipulation of the two systems for relaxed required IT skills. Besides presenting new powerful means for manipulating, exploring and extracting structured text from digitised material, the workshop gives DH researchers the opportunity to have an idea about the challenges of integrating such tools in digitisation pipelines. This includes providing background on the different factors interfering in the performance of machine learning models, such as information modelling, text encoding, OCR system choice and annotation consistency.
Figure 4: Nouveau Larousse illustré, dictionnaire universel encyclopédique, merce, de l’industrie, de la magistrature et de l’administration, 1898, p. 9

Figure 5: RDA, No 37 (august 1873).

Figure 6: Annuaire-almanach du commerce, de l’industrie, de la magistrature et de l’administration, 1871, p. 200

Figure 7: I Bienal do museu de arte moderna de São Paulo, 1951, p. 104
References


