



Text Mining with Machine Learning: Reference Extraction Using GROBID

Jacopo Notarstefano
jacopo.notarstefano [at] cern.ch



November 24th, 2015

The main features of INSPIRE are the citation extraction and their proper linking. Every year we ingest $\sim 50\text{K}$ records, and we extract from them $\sim 1.6\text{M}$ citations.

This process is facilitated by trained cataloguers, called *curators*, and currently constitutes the major part of the manual work done at INSPIRE. Clearly, this is a bottle neck.

Curators often find themselves in one of the following situations:

- 1 They only have the PDF of a paper, but no metadata.
- 2 The source provides only part of the metadata, but they want to extract more from the PDF.
- 3 They want to save the time spent entering metadata manually.

GROBID (GeneRation Of Bibliographical Data) is a machine learning library for parsing unstructured PDFs in structured XML documents, with a focus on technical and scientific publications.

It is a Java library that wraps Wapiti, a C++ toolkit for segmenting and labeling sequences using Conditional Random Fields. Its starting point is the output of `pdftoxml`, which retains much more of the PDF structure than `pdftotext`.

GROBID has been adapted to the HEP use case by Joseph Boyd under the supervision of Gilles Louppe. GROBID is widely used, for example at ResearchGate, Mendeley, HAL...

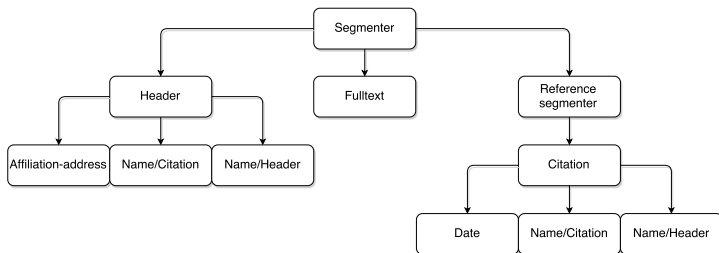
We use Conditional Random Fields because extracting metadata from a PDF can be reduced to labeling a sequence. Let's consider for example a reference:

G. Isidori and F. Teubert, *Status of indirect searches for New Physics with heavy flavour decays after the initial LHC run*, *Eur.Phys.J.Plus* **129** (2014) 40

The task of extracting metadata can be seen as labeling each word with a category, encoded in this case as a color:

G. Isidori and F. Teubert, *Status of indirect searches for New Physics with heavy flavour decays after the initial LHC run*, *Eur.Phys.J.Plus* **129** (2014) 40

GROBID is nothing more than a cascade of models, each acting on the output of the previous one.



GROBID's output: TEI

TEI (Text Encoding Initiative) publishes a set of guidelines¹ which specify encoding methods for machine-readable texts. By extension, we will call “TEI” the format described by these guidelines. GROBID's output conforms to a subset of TEI.

```
<biblStruct xml:id="b2">
  <analytic>
    <title level="a" type="main">
      Status of indirect searches for New Physics with heavy flavour decays after the initial LHC run
    </title>
    <author>
      <persName><forename type="first">G</forename><surname>Isidori</surname></persName>
    </author>
    <author>
      <persName><forename type="first">F</forename><surname>Teubert</surname></persName>
    </author>
  </analytic>
  <monogr>
    <title level="j">Eur.Phys.J.Plus</title>
    <imprint>
      <biblScope unit="volume">129</biblScope>
      <biblScope unit="issue">40</biblScope>
      <date type="published" when="2014" />
    </imprint>
  </monogr>
</biblStruct>
```

¹<http://www.tei-c.org/Guidelines/>

“Converting PDF to XML is a bit like
converting hamburgers into cows.”

— Michael Kay

That is, GROBID is not magic: it will misclassify things in various ways, and requires lots of training data to function properly.

DEMO