



Machine Learning Reference Extraction Using GROBID

Jacopo Notarstefano
jacopo.notarstefano [at] cern.ch



October 14th, 2015

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 in BibEdit.

GROBID could help them solve these problems.

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.

GROBID is a Java library that wraps Wapiti, a C++ toolkit for segmenting and labeling sequences.

GROBID is widely used, for example at ResearchGate, Mendeley, HAL...

Let's see why extracting metadata from a PDF can be reduced to labeling a sequence. Let's consider for example a reference:

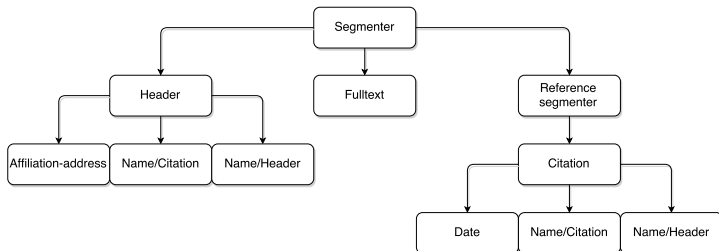
G. Isidori and F. Teubert, *Status of indirect searches for New Physics with heavy flavour decays after the initial LHC run*, *Eur.Phys.J.Plus* **129** (2014) 40

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 architecture

In GROBID parlance, the *knowledge* that is used to extract metadata from data is called a **model**. 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 guidelines 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>
```

GROBID is the work of Patrice Lopez, developer at INRIA. It has been adapted to papers from the HEP community by Joseph Boyd as part of his Master's Thesis for EPFL, under the supervision of Gilles Louppe, Senior Fellow at Inspire.

In particular, Joseph added to GROBID the concept of a *collaboration*, such as ATLAS or CMS, and improved considerably the accuracy of GROBID's models by providing lots of training data, taken from Inspire.

“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.

Unfortunately, the XML output is not compatible with what we use in Inspire. So we need to add some layer between GROBID and Invenio: `invenio-grobid`.

`invenio-grobid` is the joint work of me, my supervisor Jan Åge Lavik and Ilias Koutsakis.

`invenio-grobid` has been recently released on PyPI and is available at <https://github.com/inspirehep/invenio-grobid>.

DEMO

Current status and tentative roadmap, 1/2

- 1
 - Allow catalogers to upload PDFs
 - Automatic extraction of metadata
 - Possibility to push to system
- 2
 - Allow catalogers to edit extracted metadata
 - Perhaps integrate JSONEditor?

Current status and tentative roadmap, 2/2

- 3
 - Take over bibliographic reference parsing (i.e. kill Refextract)
 - Consolidate with other services (e.g. CrossRef)
 - Integrate in automatic ingestion workflows
- 4
 - Pre-fill user submission forms from PDF
- 5
 - `invenio-grobid` could be used to feed back corrections in GROBID, improving its precision
 - Advanced visual interface to correct extraction (e.g. side-by-side view)