Grobid:
Extraction of metadata

Annette Holtkamp

CERN-UNESCO School on Digital Libraries

Nairobi, Oct 8 – 12, 2018

Grobid: GeneRation Of Bibliographic Data

- Open source machine learning library
- extracts, parses, re-structures pdf documents into structured TEI-encoded documents
- focus on technical and scientific publications

Documentation: https://grobid.readthedocs.io/

Functionalities

- Header extraction
  - Title, abstract, authors, affiliations, keywords…
- Reference extraction
- Parsing of author names
- Parsing of affiliation and address blocks
- Parsing of dates
- Full text extraction from pdf
  - Structure of the text body

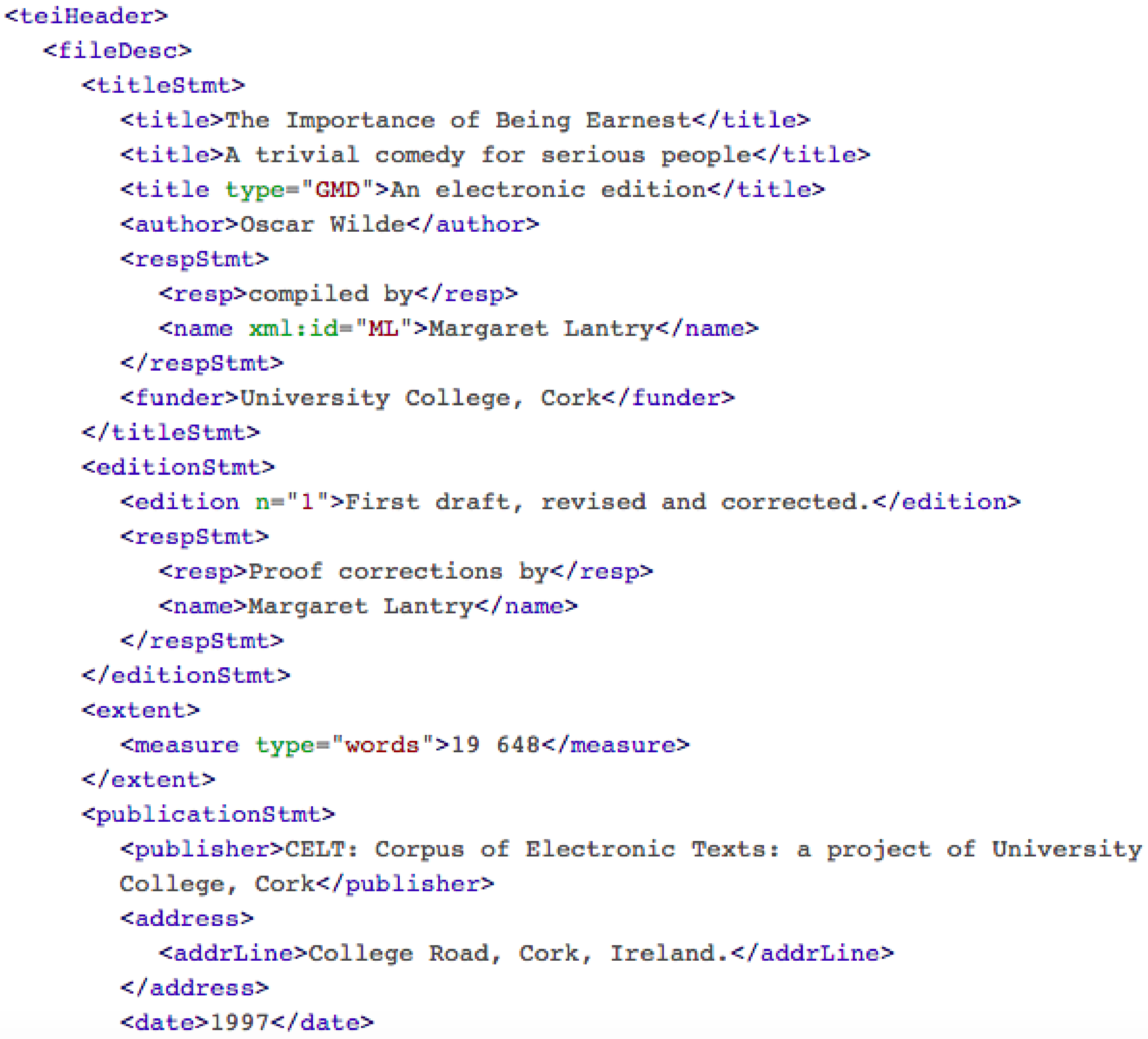
Features

- Batch processing
- API
  - RESTful, JAVA
- Semi-automatic generation of training data
- Output in TEI format

TEI

- Text Encoding Initiative
- Goal: xml standard for the representation of texts in digital form
- Guidelines specifying encoding methods for machine-readable texts
- Very precise text annotations while maintaining human readability

http://www.tei-c.org/



```
<teiHeader>
   <fileDesc>
      <titleStmt>
         <title>The Importance of Being Earnest</title>
         <title>A trivial comedy for serious people</title>
         <title type="GMD">An electronic edition</title>
         <author>Oscar Wilde</author>
         <respStmt>
            <resp>compiled by</resp>
            <name xml:id="ML">Margaret Lantry</name>
         </respStmt>
         <funder>University College, Cork</funder>
      </titleStmt>
      <editionStmt>
         <edition n="1">First draft, revised and corrected.</edition>
         <respStmt>
            <resp>Proof corrections by</resp>
            <name>Margaret Lantry</name>
         </respStmt>
      </editionStmt>
      <extent>
         <measure type="words">19 648</measure>
      </extent>
      <publicationStmt>
         <publisher>CELT: Corpus of Electronic Texts: a project of University
         College, Cork</publisher>
         <address>
            <addrLine>College Road, Cork, Ireland.</addrLine>
         </address>
         <date>1997</date>
```

Exercise

- Go to http://cloud.science-miner.com/grobid/
- Click on TEI tab
- Select article pdf on your computer
- Choose service
  - E.g. process header document
- Inspect result