DARIAH Second Code Sprint: Track A

Acknowledgment Section Processing of GROBID

Tanti Kristanti
INRIA-Paris

24-26 September 2019
Berlin

2018

- Github Repository
  https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining
- Demonstrator
  http://destracka.herokuapp.com/
- Functionalities:
  - Citation extraction of Pdf files using Grobid
  - Extraction of some additional information from external service (e.g., entity-fishing[1])
  - Visualisation of extracted information on the Pdf files

2019

The author is grateful to Dr. Ibrahim ElAgib of King Saud University, College of Sciences, Physics & Astronomy Department, for valuable discussions

$ **curl -X POST -d** "acknowledgments=The author is grateful to Dr. Ibrahim ElAgib of King Saud University, College of Sciences, Physics & Astronomy Department, for valuable discussions." localhost:8070/api/processAcknowledgments

```
<acknowledgment>
        <affiliation>King Saud University , College of Sciences</affiliation>
        <individual>Dr . Ibrahim ElAgib</individual>
</acknowledgment>
```

[1] entity-fishing, 2019, https://github.com/kermitt2/entity-fishing

A Glimpse of GROBID

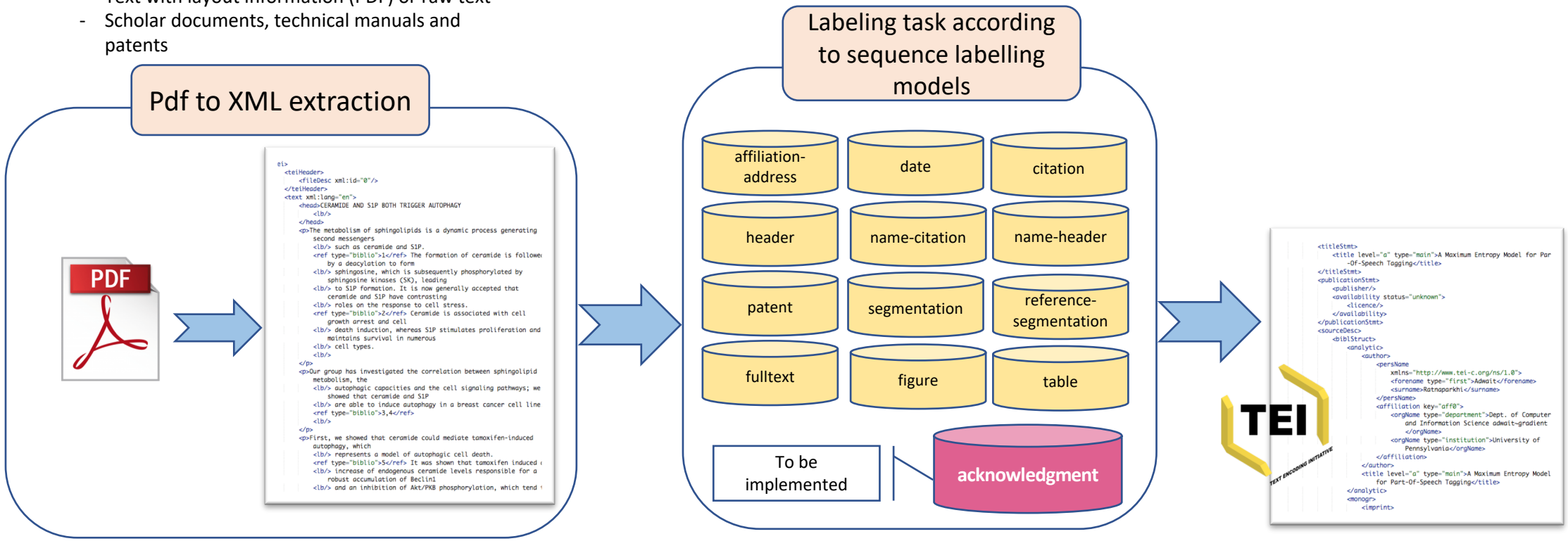
- **G**ene**R**ation **O**f **BI**bliograhic **D**ata[1)]
  - Started as a hobby in 2008
  - Available in open source since 2011
  - Distributed under Apache 2.0 license
- A machine learning library for extracting, parsing and re-structuring raw documents into structred XML/TEI encoded documents.
  - Focus on technical and scientific publications
- Batch processing, RESTful API, Java API, docker container
- Git source : https://github.com/kermitt2/grobid
- Documentation: http://grobid.readthedocs.org/

[1] Patrice Lopez, 2019, https://github.com/kermitt2/grobid

Grobid's Architecture (High-level Segmentation)

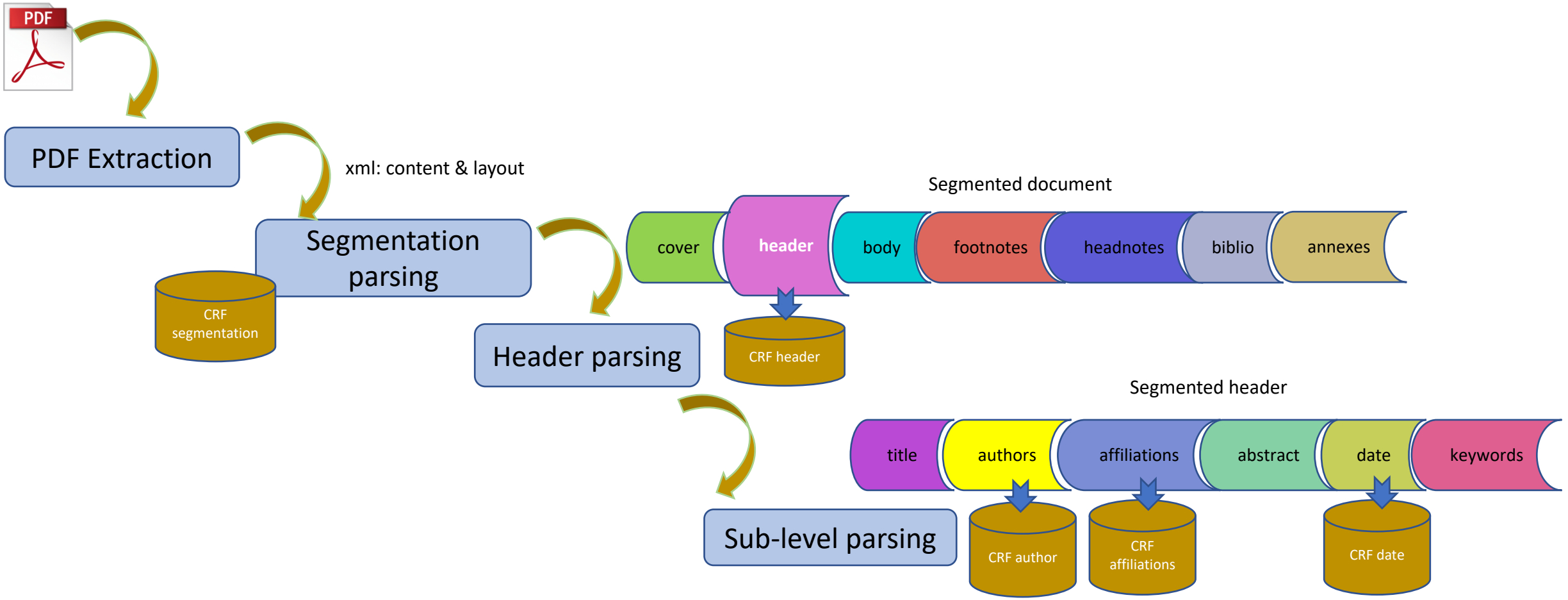
Input: [1)]
- Technical and scientific domains
- Text with layout information (PDF) or raw text
- Scholar documents, technical manuals and patents

Machine learning approach: cascading of linear chain CRF.
For a complex extraction and parsing tasks, several models are used in **cascade**.



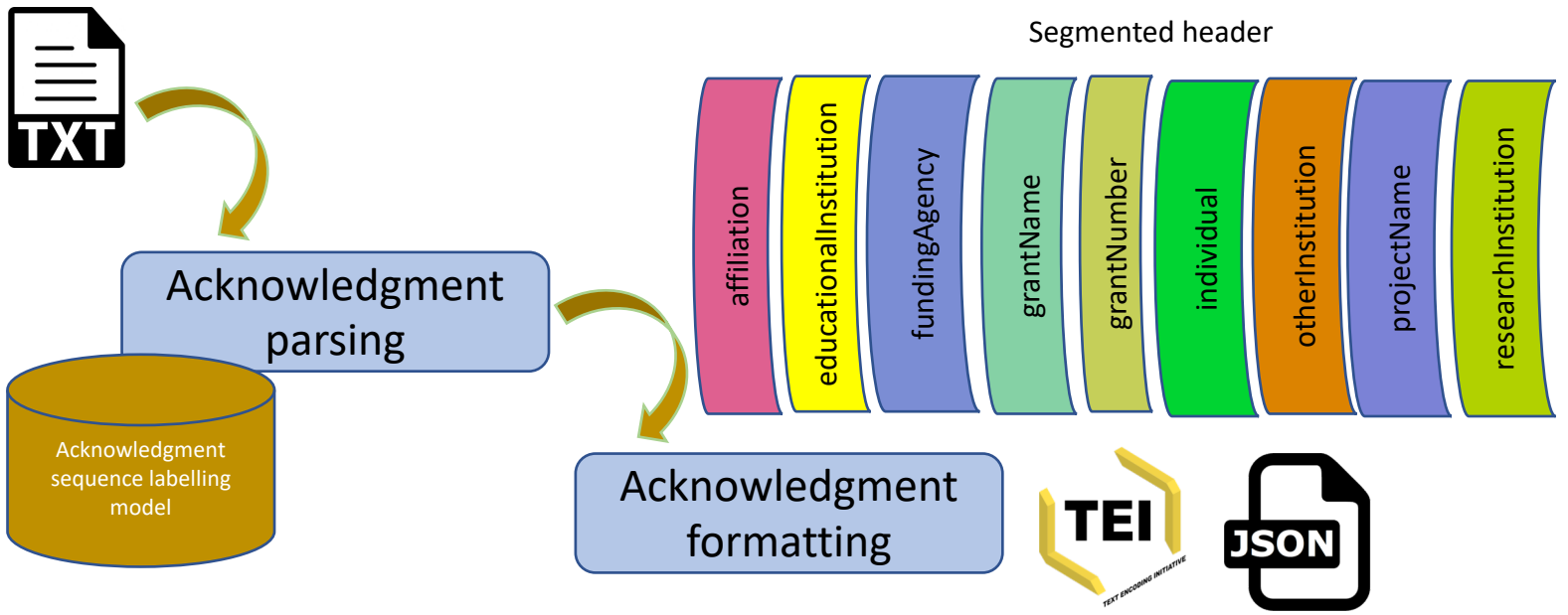
[1] Patrice Lopez, April 2015, https://grobid.readthedocs.io/en/latest/grobid-04-2015.pdf
[2] CRF++: Yet Another CRF toolkit, https://taku910.github.io/crfpp/#templ

Header Processing with Grobid



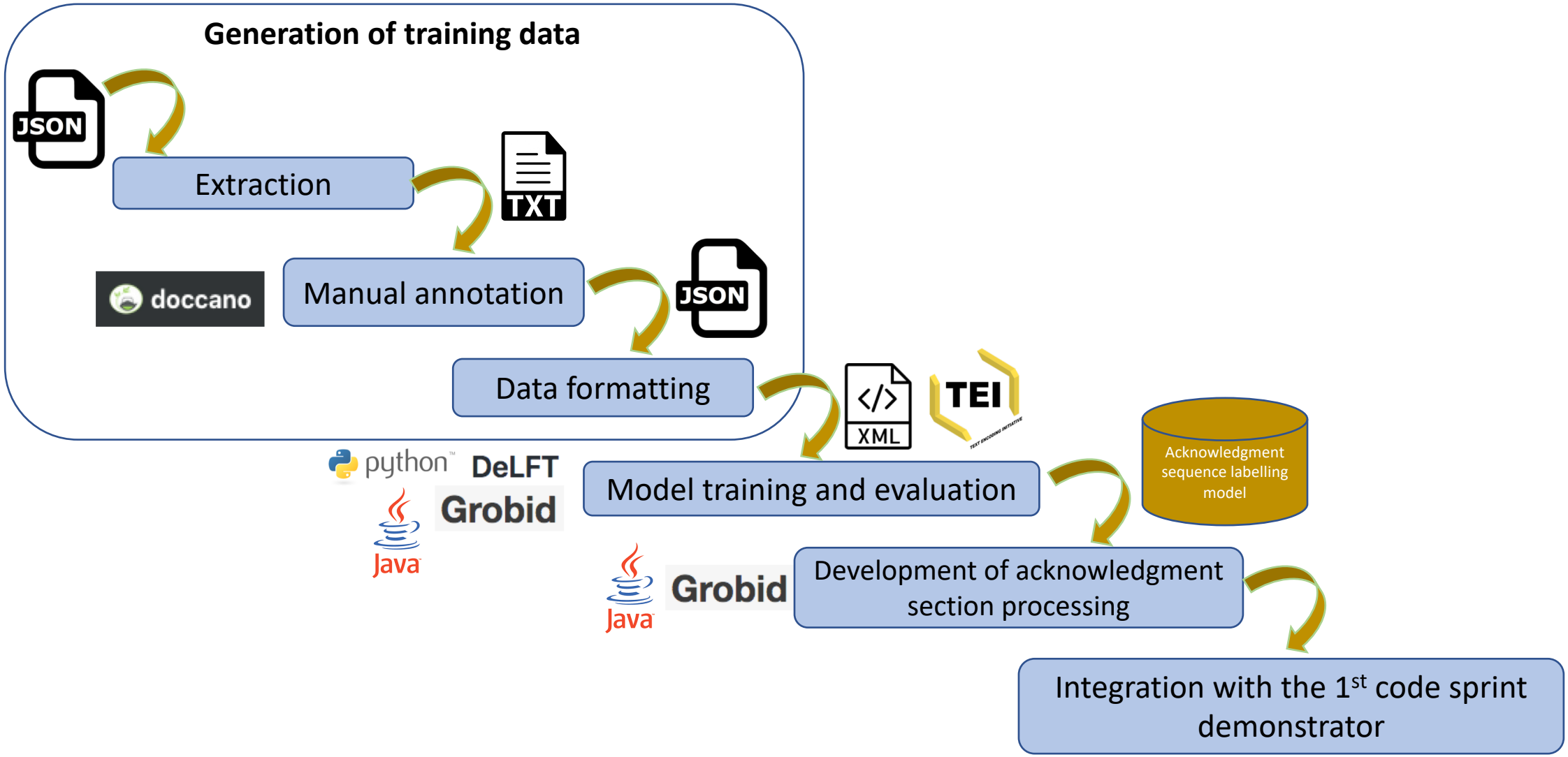
[1] Patrice Lopez, April 2015, https://grobid.readthedocs.io/en/latest/grobid-04-2015.pdf

Acknowledgment Section Processing (to-be)



[1] TEI Consortium, July 2019, https://tei-c.org/release/doc/tei-p5-doc/en/html/SG.html

How?


Generation of training data
JSON
Extraction
TXT
doccano
Manual annotation
JSON
Data formatting
XML
TEI
python
DeLFT
Grobid
Java
Model training and evaluation
Acknowledgment sequence labelling model
Java
Grobid
Development of acknowledgment section processing
Integration with the 1st code sprint demonstrator

Activities

- Day 1
  - A. Download and installation
    - Track A of DARIAH ERIC Repo (https://github.com/DARIAH-ERIC/DESIR-CodeSprint-TrackA-TextMining)
    - Doccano (https://github.com/chakki-works/doccano)
    - DeLFT (https://github.com/kermitt2/delft)
      - Including pre-trained word embeddings, e.g. *glove Common Crawl* (nlp.stanford.edu/data/glove.840B.300d.zip)
    - Grobid (https://github.com/kermitt2/grobid)
  - B. Annotations with Doccano
- Day 2
  - Continuing the annotation activities
  - Preprocessing the datasets file to be read by DeLFT and Grobid
  - Training and evaluation with DeLFT
  - Create a model with Grobid
- Day 3
  - Continue to make models with Grobid
  - Create acknowledgment API service for text
  - Create acknowledgment API service for Pdf files (optional)
  - Call the service to Demonstrator Track A (optional)

DeLFT: Clone, Build and Run

- Clone:

  $ git clone https://github.com/kermitt2/delft.git

- Setup first a virtual environment

  $ virtualenv --system-site-packages -p python3 env

  $ source env/bin/activate

- Install dependencies:

  $ pip3 install -r requirements.txt

- Download some pre-trained word embeddings

  glove Common Crawl

[1] A Deep Learning Framework for Text (DeLFT), 2019, https://github.com/kermitt2/delft

DeLFT Dependencies

- keras== 2.2.4
- numpy== 1.16.1
- pandas== 0.24.2
- bleach>=2.1.0
- regex==2018.2.21
- scikit-learn== 0.20.3
- tqdm>=4.21
- tensorflow>=1.12.0
  - if the computer has GPU, **tensorflow_gpu**==1.12.0
- gensim==3.4.0
- langdetect==1.0.7
- textblob==0.15.1
- h5py==2.7.1
- unidecode==1.0.22
- pydot==1.2.4
- lmdb==0.94
- keras-bert>=0.39.0

- Example of retraining the CoNLL 2003 model:
  - Put the datasets under data/sequenceLabelling/CoNLL-2003
  - Copy the model for example under **embeddings/**
    - glove Common Crawl : http://nlp.stanford.edu/data/glove.840B.300d.zip
  - Change the path to the model in **embedding-registry.json**
  - Train and evaluate
  - $ python3 nerTagger.py --dataset-type conll2003 train_eval

```
"embeddings": [
    {
        "name": "glove-840B",
        "path": "embeddings/glove.840B.300d.txt",
        "type": "glove",
        "format": "vec",
        "lang": "en",
        "item": "word"
    },
```

GROBID: Clone, Build and Run

- Clone:

  $ git clone https://github.com/kermitt2/grobid.git

- $ cd grobid
- Build:
  - ./gradlew clean install
- Run the server (on the default port 8070):
  - ./gradlew run

[1] GROBID, 2019, https://grobid.readthedocs.io/en/latest/Grobid-service/

Pre-processing of Training Data

- Doccano can read CSV or JSON file. Format of JSON file :
  - Each line contains a JSON object with a text key while other information (identifier, title) as « metadata »
- Delft and Grobid read TEI-XML file
- File in Json and TEI-XML format : TrackA-TextMining/data/secondCodeSprint
- Codes: TrackA-TextMining/src/main/java/org/dariah/desir/secondeCodeSprint

Doccano


```
{"meta":{"identifier":"76b3ed4f-086b-41cf-a693-d59f09a6a394","title":"Generated Covariates in Nonparametric Estimation: A Short Review"},"text":"This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 \"Economic Risk\". http:\/\/sfb649.wiwi.hu-berlin.de ISSN 1860-5664"}
{"meta":{"identifier":"76614818-943f-48e6-95d7-e63045074c75","title":"Effect of a single injection of gonadotropin-releasing hormone (GnRH) and human chorionic gonadotropin (hCG) on testicular blood flow measured by color doppler ultrasonography in male Shiba goats"},"text":"ACKNOWLEDGMENT. We are grateful to Dr. G.D. Niswender (Animal Reproduction and Biotechnology Laboratory, Colorado State University, Fort Collins, CO, U.S.A.) for providing antisera to estradiol-17β (GDN 244) and testosterone (GDN 250)."}
{"meta":{"identifier":"76e4bdd3-867a-4732-b349-1729ad668707","title":"Elucidating a normal function of huntingtin by functional and microarray analysis of huntingtin-null mouse embryonic fibroblasts"},"text":"We are grateful to the personal of UTSW Microarray Core for assistance with these experiments and to Janet Young for administrative assistance. We are thankful to Thomas Südhof and Katsuhiko Tabuchi for a generous gift of Lenti-NLS-GFP-Cre virus. This study was supported by the Hereditary Disease Foundation and NINDS R01 NS38082 and R01 NS056224 (IB), NINDS R01 NS043466 (SZ), the NIH GM59419 (GH), and the Ara Parseghian Medical Research Foundation (JR)."}
{"meta":{"identifier":"76e72e3e-1089-447c-bba6-fd3ec1b57104","title":"Genetic alterations of WWOX in Wilms\u2019 tumor are involved in its carcinogenesis"},"text":"The authors wish to thank Ewa Latkowska for her excellent technical support. This study was supported by the Polish Ministry of Science and Higher Education 2670\/B\/P01\/2008\/34 and 0757\/B\/P01\/2009\/37. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript."}
{"meta":{"identifier":"76bb6553-6c8f-4a4a-bae8-8d1c5cde336e","title":"Using Semantic Web technologies for the generation of domain-specific templates to support clinical study metadata standards"},"text":"The authors thank Dr. Chunhua Weng from Columbia University and Dr. Cui Tao from Mayo Clinic who participated in the evaluation. The authors also thank the technical support from Mr. Craig Stancl from Mayo Clinic. The study is supported in part by the SHARP Area 4: Secondary Use of EHR Data (90TR000201)."}
```


DeLFT & GROBID

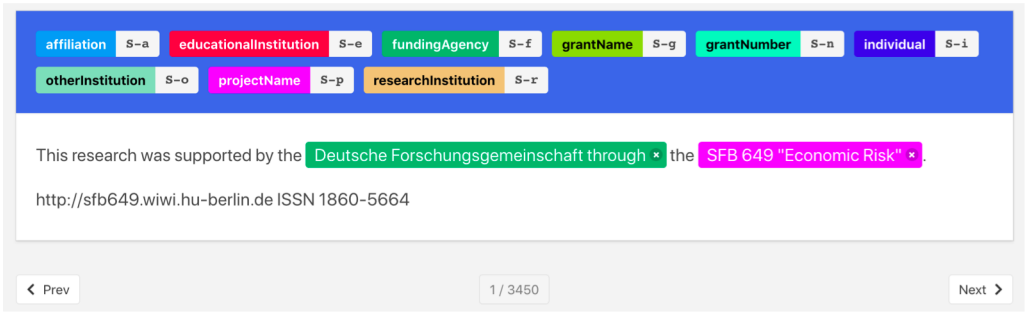

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<TEI
    xmlns:mml="http://www.w3.org/1998/Math/MathML"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns="http://www.tei-c.org/ns/1.0">
    <acknowledgments>
        <p type="acknowledgment">This research was supported by
            <rs type="fundingAgency">the Deutsche Forschungsgemeinschaft</rs>
            through
            <rs type="projectName">the SFB 649 "Economic Risk"</rs>. http
            ://sfb649.wiwi.hu-berlin.de ISSN 1860-5664
        </p>
        <p type="acknowledgment">ACKNOWLEDGMENT. We are grateful to
            <rs type="individual">Dr. G.D. Niswender</rs> (
            <rs type="affiliation">Animal Reproduction and Biotechnology
            Laboratory, Colorado State University, Fort Collins, CO, U.S.A
            .</rs>) for providing antisera to estradiol-17β (GDN 244) and
            testosterone (GDN 250).
        </p>
        <p type="acknowledgment">We are grateful to the personal of
            <rs type="affiliation">UTSW Microarray Core</rs> for assistance with
            these experiments and to
            <rs type="individual">Janet Young</rs> for administrative assistance
            We are thankful to
            <rs type="individual">Thomas Südhof</rs> and
            <rs type="individual">Katsuhiko Tabuchi</rs> for a generous gift of
            Lenti-NLS-GFP-Cre virus. This study was supported by
            <rs type="fundingAgency">the Hereditary Disease Foundation</rs> and
            <rs type="fundingAgency">NINDS R01 NS38082</rs> and
            <rs type="fundingAgency">R01 NS056224 (IB)</rs>,
            <rs type="fundingAgency">NINDS R01 NS043466 (SZ)</rs>,
            <rs type="fundingAgency">the NIH GM59419 (GH)</rs>, and
            <rs type="fundingAgency">the Ara Parseghian Medical Research
            Foundation (JR)</rs>.
        </p>
```



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<TEI
    xmlns:mml="http://www.w3.org/1998/Math/MathML"
    xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns="http://www.tei-c.org/ns/1.0">
    <acknowledgments>
        <acknowledgment>This research was supported by
            <fundingAgency>the Deutsche Forschungsgemeinschaft</fundingAgency>
            through
            <projectName>the SFB 649 "Economic Risk"</projectName>. http://sfb649
            .wiwi.hu-berlin.de ISSN 1860-5664
        </acknowledgment>
        <acknowledgment>ACKNOWLEDGMENT. We are grateful to
            <individual>Dr. G.D. Niswender</individual> (
            <affiliation>Animal Reproduction and Biotechnology Laboratory,
            Colorado State University, Fort Collins, CO, U.S.A.</affiliation
            >) for providing antisera to estradiol-17β (GDN 244) and
            testosterone (GDN 250).
        </acknowledgment>
        <acknowledgment>We are grateful to the personal of
            <affiliation>UTSW Microarray Core</affiliation> for assistance with
            these experiments and to
            <individual>Janet Young</individual> for administrative assistance.
            We are thankful to
            <individual>Thomas Südhof</individual> and
            <individual>Katsuhiko Tabuchi</individual> for a generous gift of
            Lenti-NLS-GFP-Cre virus. This study was supported by
            <fundingAgency>the Hereditary Disease Foundation</fundingAgency> and
            <fundingAgency>NINDS R01 NS38082</fundingAgency> and
            <fundingAgency>R01 NS056224 (IB)</fundingAgency>,
            <fundingAgency>NINDS R01 NS043466 (SZ)</fundingAgency>,
            <fundingAgency>the NIH GM59419 (GH)</fundingAgency>, and
            <fundingAgency>the Ara Parseghian Medical Research Foundation (JR
            )</fundingAgency>.
        </acknowledgment>
        <acknowledgment>The authors wish to thank
            <individual>Ewa Latkowska</individual> for her excellent technical
            support. This study was supported by
            <fundingAgency>the Polish Ministry of Science and Higher Education
            </fundingAgency>
            <grantNumber>2670/B/P01/2008/34</grantNumber> and
            <grantNumber>0757/B/P01/2009/37</grantNumber>. The funders had no
            role in study design, data collection and analysis, decision to
            publish, or preparation of the manuscript.
        </acknowledgment>
```

Annotation with Doccano

- Source : https://github.com/chakki-works/doccano
- Git clone : https://github.com/chakki-works/doccano.git
- Installation (there are 3 options):
  - Pull the production docker image
    - $ docker pull chakkiworks/doccano
- Usage (there are 3 options):
  - Running the docker image as a container
    - $ docker run -d --name doccano -p 8000:8000 chakkiworks/doccano
    - $ docker exec doccano tools/create-admin.sh "admin" "admin@example.com" "password"
    - Remove existing docker image: $ docker rm doccano
- Open in http://127.0.0.1:8000/login/
- Sequence labeling : https://github.com/chakki-works/doccano/wiki/A-Tutorial-For-Sequence-Labeling-Project
- Data to be annotated:
  TrackA-TextMining/data/secondCodeSprint/json/raw/acknowledgementsFormattedForDoccano.json
- Export the annotated data (JSONL format)



Inter Annotator Agreement

- Reconciliation (e.g. “the”)
- Decision in case of ambiguity
- Documentation in https://grobid.readthedocs.io/en/latest/training/General-principles/

Acknowledgment Sequence Labelling Model with DeLFT

- Datasets:
  - Train: data/sequenceLabelling/desirSecondCodeSprint/train.xml
  - Validation: data/sequenceLabelling/desirSecondCodeSprint/valid.xml
- Model built : data/models/sequenceLabelling/acknowledgment
- Acknowledgment tagger : acknowledgmentTagger.py
  - « train » for training
    - $ python3 acknowledgmentTagger.py train
  - « annotate » for labelling a text and returing a list of acknowledgment parsing results
    - $ python3 acknowledgmentTagger.py tag

```
Epoch 1/50
9/9 [==============================] - 13s 1s/step - loss: 552.3135
    f1 (micro): 0.00
Epoch 2/50
9/9 [==============================] - 8s 931ms/step - loss: 180.8875
    f1 (micro): 1.65
Epoch 3/50
9/9 [==============================] - 9s 967ms/step - loss: 85.6009
    f1 (micro): 4.08
...

Epoch 46/50
9/9 [==============================
    f1 (micro): 50.92
Epoch 47/50
9/9 [==============================
    f1 (micro): 51.02
```

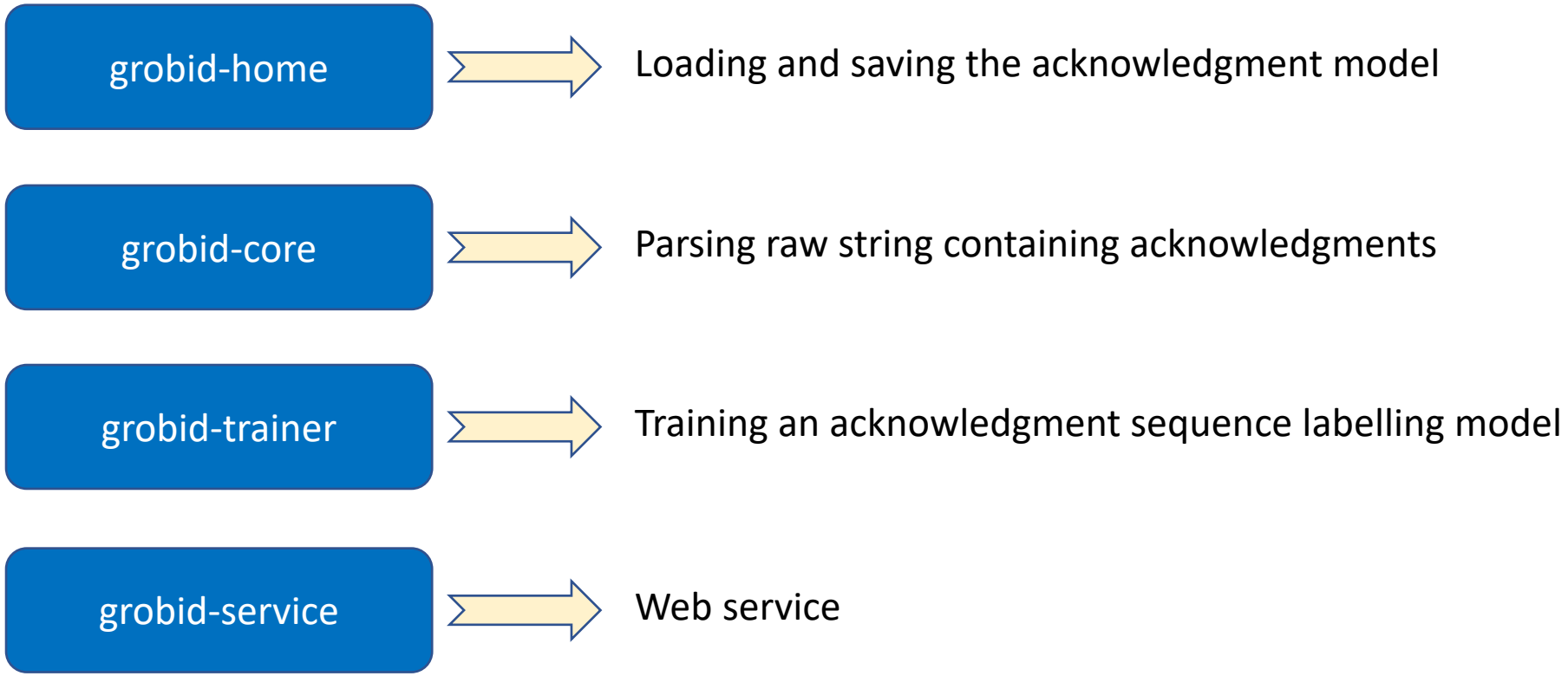

```
"software": "DeLFT",
"date": "2019-08-21T16:32:31.673162",
"model": "acknowledgment",
"texts": [
  {
    "text": "We want to particularly acknowledge the patients enrolled in this
      study for their participation and the Basque Biobank for Research-OEHUN for
      its collaboration providing the human samples and the clinical information
      used in this project with appropriate ethics approval. Our gratefulness to
      Dr. Juan Burgos for the selection of the human samples and Dr. Felix Royo
      for helping with statistical analysis.",
    "entities": [
      {
        "text": "Basque Biobank for Research - OEHUN",
        "class": "<fundingAgency>",
        "score": 1,
        "beginOffset": 104,
        "endOffset": 136
      },
      {
        "text": "Dr . Juan Burgos",
        "class": "<individual>",
        "score": 1,
        "beginOffset": 292,
        "endOffset": 306
      },
      {
        "text": "Dr . Felix Royo",
        "class": "<individual>",
        "score": 1,
        "beginOffset": 351,
        "endOffset": 364
      }
    ]
  }
],
"runtime": 0.438
```

Acknowledgment Sequence Labelling Model with GROBID

- Model built: grobid-home/models/acknowledgment
- Datasets (80:20) : grobid-trainer/resources/dataset/acknowledgment/corpus (for training) and grobid-trainer/resources/dataset/acknowledgment/evaluation (for evaluation)
- CRFPP-templates[1)] can be copied according to the model used, for instance from : grobid-trainer/resources/dataset/affiliation-address/crfpp-templates
- Temporary training and testing files : grobid-home/tmp
- Classes to be developed are in 3 sub-module directories :
  - grobid-core
  - grobid-training
  - grobid-service

[1] taku@chasen.org, CRF++: Yet Another CRF toolkit, 06/01/2003, https://taku910.github.io/crfpp/#templ

Sub Modules


grobid-home
Loading and saving the acknowledgment model
grobid-core
Parsing raw string containing acknowledgments
grobid-trainer
Training an acknowledgment sequence labelling model
grobid-service
Web service

Acknowledgment Model Training and Evaluation with Grobid

- $ ./gradlew train_acknowledgment

```
sourcePathLabel: /Users/tkristan/Desktop/Tanti/grobid/grobid-trainer/../grobid
    -home/../grobid-trainer/resources/dataset/acknowledgment/corpus
outputPath for training data: /Users/tkristan/Desktop/Tanti/grobid/grobid-home
    /tmp/acknowledgment5013200036500975790.train
1 tei files
        epsilon: 1.0E-5
        window: 20
        nb max iterations: 2000
        nb threads: 8
* Load patterns
* Load training data
* Initialize the model
* Summary
    nb train:    176
    nb labels:   18
    nb blocks:   74940
    nb features: 1349226
* Train the model with l-bfgs
  [   1] obj=25173.59   act=137531   err=48.69%/100.00% time=0.49s/0.49s
  [   2] obj=19685.45   act=158148   err=48.69%/100.00% time=0.33s/0.82s
  ...
  [ 286] obj=908.52     act=4808     err= 0.12%/ 2.27% time=0.35s/108.21s
  [ 287] obj=908.24     act=4767     err= 0.12%/ 2.27% time=0.35s/108.56s
* Save the model
* Done
Model for acknowledgment created in 109268 ms
sourcePathLabel: /Users/tkristan/Desktop/Tanti/grobid/grobid-trainer/../grobid
    -home/../grobid-trainer/resources/dataset/acknowledgment/evaluation
outputPath for training data: /Users/tkristan/Desktop/Tanti/grobid/grobid-home
    /tmp/acknowledgment7540005405272076 99.test
1 tei files
Aug 19, 2019 2:04:46 PM org.grobid.core.jni.WapitiModel init
INFO: Loading model: /Users/tkristan/Desktop/Tanti/grobid/grobid-trainer
    /../grobid-home/models/acknowledgment/model.wapiti (size: 1354737)
[Wapiti] Loading model: "/Users/tkristan/Desktop/Tanti/grobid/grobid-trainer
    /../grobid-home/models/acknowledgment/model.wapiti"
Model path: /Users/tkristan/Desktop/Tanti/grobid/grobid-trainer/../grobid-home
    /models/acknowledgment/model.wapiti
Labeling took: 114 ms
```

[1] GROBID, Training and evaluating GROBID models, 2019, https://grobid.readthedocs.io/en/latest/Training-the-models-of-Grobid/

```
===== Token-level results =====

label                       accuracy     precision    recall       f1

<affiliation>               96.72        24.43        76.19        36.99
<educationalInstitution>    97.81        0            0            0
<fundingAgency>             94.44        60.88        79.92        69.12
<grantName>                 98.83        0            0            0
<grantNumber>               98.98        87.5         68.29        76.71
<individual>                97.03        83.71        84.52        84.11
<otherInstitution>          98.68        13.33        18.18        15.38
<projectName>               99.52        66.67        31.58        42.86
<researchInstitution>       98.98        55.56        75           63.83

all fields                  97.89        63.44        67.38        65.3
                                                                   (micro average)
                            97.89        43.56        48.19        43.22
                                                                   (macro average)

===== Field-level results =====

label                 accuracy     precision    recall       f1

<affiliation>               97.6         37.5         54.55        44.44
<educationalInstitution>    98.72        0            0            0
<fundingAgency>             89.78        37.5         34.62        36
<grantName>                 98.88        0            0            0
<grantNumber>               96.49        66.67        48.28        56
<individual>                93.77        76.34        80.68        78.45
<otherInstitution>          97.76        14.29        11.11        12.5
<projectName>               99.52        50           33.33        40
<researchInstitution>       98.56        28.57        33.33        30.77

all fields                  96.79        58.25        53.05        55.53
                                                                   (micro average)
                            96.79        34.54        32.88        33.13
                                                                   (macro average)

===== Instance-level results =====

Total expected instances:   41
Correct instances:          7
Instance-level recall:      17.07

Evaluation for acknowledgment model is realized in 362 ms
```

GROBID Acknowledgment Section Processing Result

$ **curl -X POST -d** "acknowledgments=The author is grateful to Dr. Ibrahim ElAgib of King Saud University, College of Sciences, Physics & Astronomy Department, for valuable discussions."
localhost:8070/api/processAcknowledgments

```
<acknowledgment>
        <affiliation>King Saud University , College of Sciences</affiliation>
        <individual>Dr . Ibrahim ElAgib</individual>
</acknowledgment>
```

Coordinates structure in the original PDF

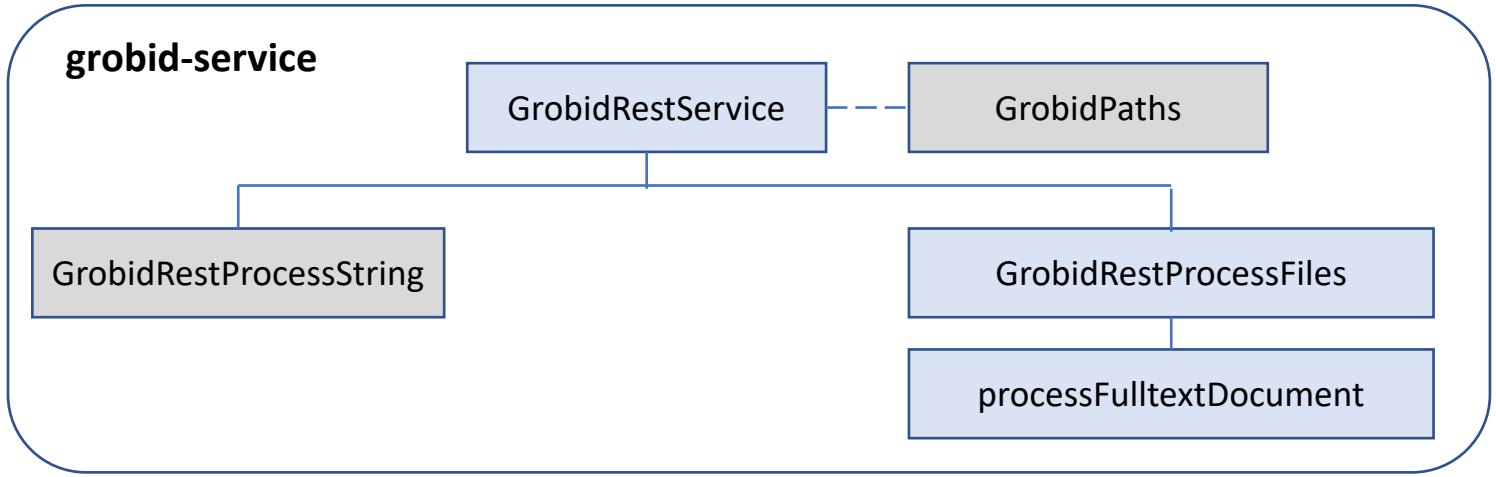
- Grobid version 0.4.2 and higher provides coordinate areas in the following document sub-structures:
  - *persName* (for a complete author name)
  - *figure* (for figure and table)
  - *ref* (for bibliographical, figure, table, and formula reference markers)
  - *biblStruct* (for a bibliographical reference)
  - *formula* (for mathematical equations)
- Coordinates are available in:
  - Full text processing → returning a TEI document
  - PDF annotation services → returning JSON

[1] GROBID, Coordinates of structures in the original PDF, 27/08/2019, https://grobid.readthedocs.io/en/latest/Coordinates-in-PDF/

Example of demanding coordinates

- Full text extraction and add coordinates to the figures (and tables) only:

  $ curl -v --form input=@./Document1.pdf --form teiCoordinates=figure --form teiCoordinates=biblStruct localhost:8070/api/processFulltextDocument




```
<div type="references">
    <listBibl>
        <biblStruct coords="14,335.40,89.54,211.08,8.07;14,335
            .40,100.46,106.58,8.07" xml:id="b0">
            <analytic>
                <title/>
                <idno>ar- ticle 8 of 18</idno>
            </analytic>
            <monogr>
                <title level="j">Anonymous. Runtime process
                    infection. Phrack</title>
                <imprint>
                    <biblScope unit="volume">11</biblScope>
                    <biblScope unit="issue">59</biblScope>
                    <date type="published" when="2002-12" />
                </imprint>
            </monogr>
        </biblStruct>
        <biblStruct coords="14,335.40,110.18,211.12,8.07;14,335
            .40,121.10,71.48,8.07" xml:id="b1">
            <analytic>
                <title/>
                <idno>article 10 of 15</idno>
            </analytic>
            <monogr>
                <title level="j">Apk. Interface promiscuity
                    obscurity. Phrack</title>
                <imprint>
                    <biblScope unit="volume">8</biblScope>
                    <biblScope unit="issue">53</biblScope>
                    <date type="published" when="1998-07" />
                </imprint>
            </monogr>
        </biblStruct>
```