JDEV2017

CRAWLING, HARVESTING, SCRAPING

Cartographie et synthèse de réseaux de collaborations scientifiques

Sonia Guérin-Hamdi, ISH CNRS

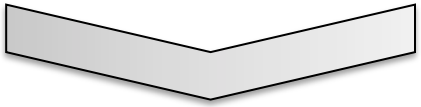
INSTITUT DES SCIENCES DE L'HOMME

Crawling, Harvesting, Scraping

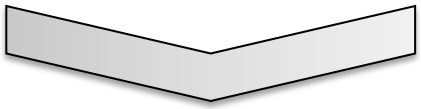
Services Web, Flux
API Rest,
RSS, OAI-PMH, Base de données SQL

Fichiers structurés
Bibliothèque Zotero,
Dump SQL, XML, oai dc,
TEI, CSV, JSON

Fichiers non structurés
PDF, TXT…



Indexation des informations structurées (métadonnées)



INDEX

Programme

1. **Comment extraire des relations de co-écriture et citations explicites issues des bibliographies?**

✓ Cas d'application avec GROBID

2. **Quelles sont les méthodes et techniques pour la constitution d'un corpus?**

✓ Moissonnage pour la constitution d'un corpus hétérogènes et dispersés.

✓ Indexation massive des données

3. **Comment synthétiser les réseaux d'auteurs et les coopérations scientifiques?**

✓ Recherche et visualisation.

1. GROBID : GeneRation Of Bibliographic Data

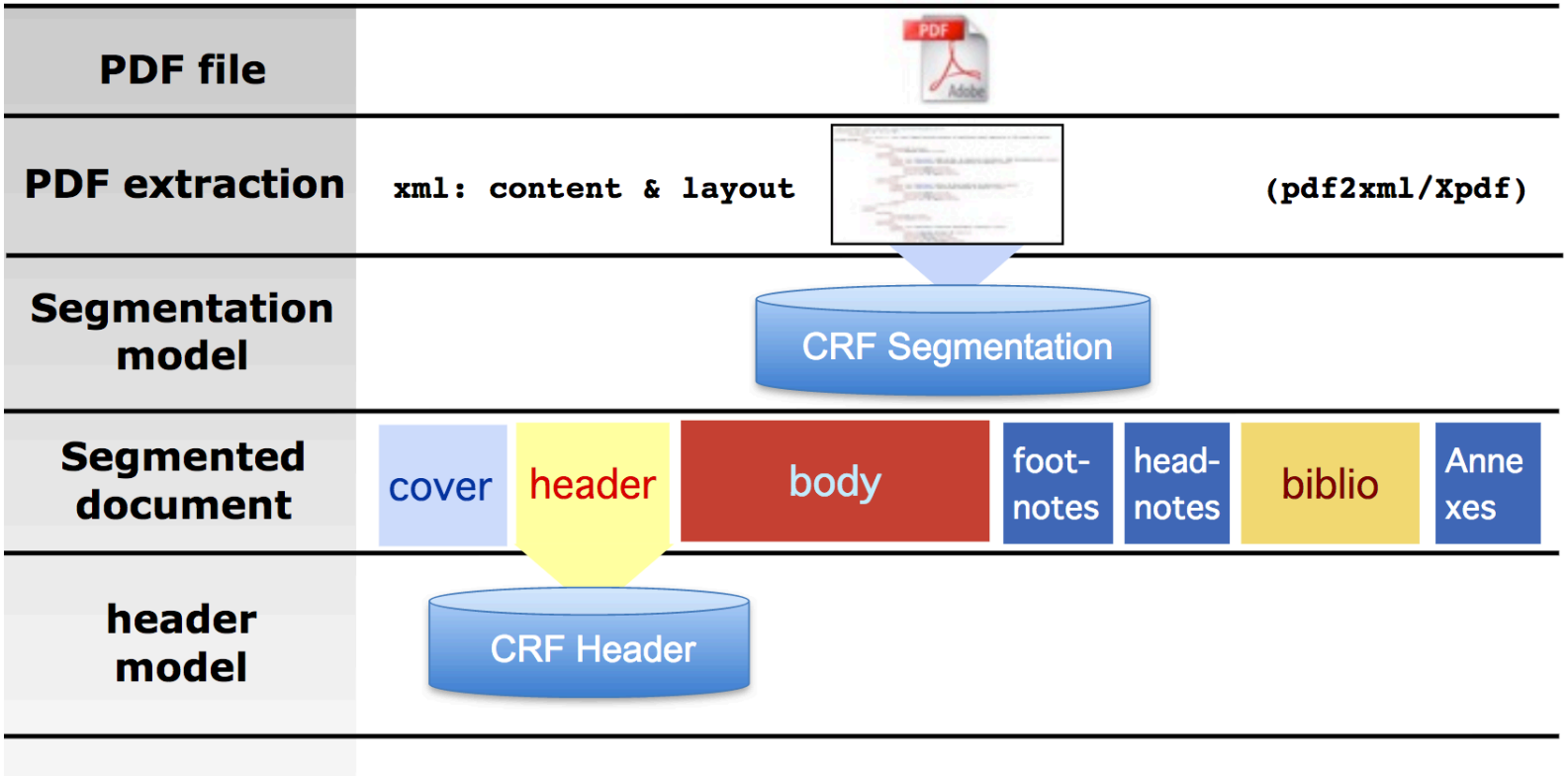
- **Objectif** : Méthodes d'extraction automatique de contenu structuré à partir de PDF.
- Analyse PDF avec CRF exploitant à la fois les fonctionnalités de mise en page et les contenus
- Approche par apprentissage automatique: cascade de modèles CRF (Conditional Random Fields)
- Sortie: **TEI XML** *Text Encoding Initiative XML*

⇒3 documents PDF par seconde,

⇒3000 références en moins de 10 seconds.

1. GROBID : GeneRation Of Bibliographic Data

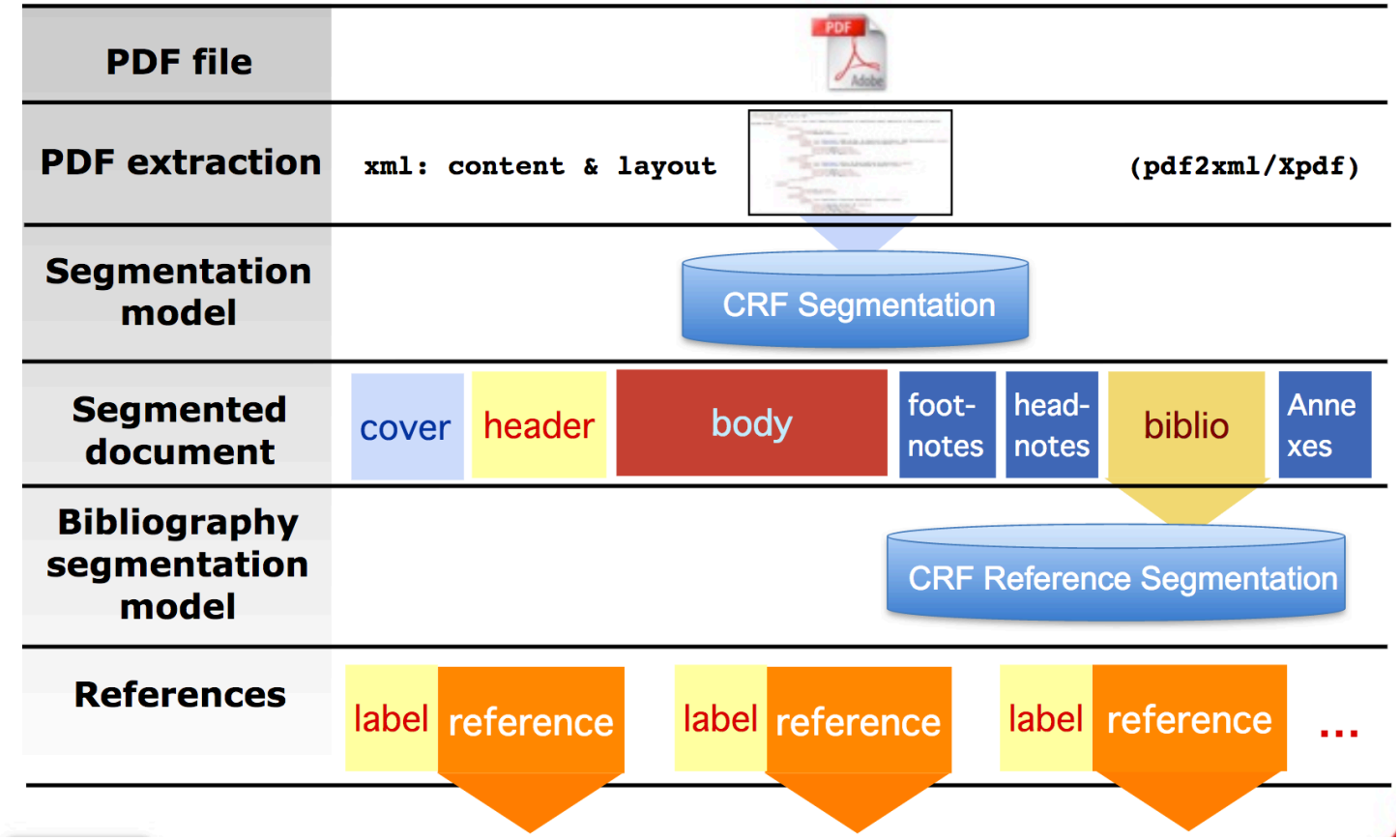
Header processing



Lopez Patrice, Inria, *GROBID : from PDF to structured documents,* Avril 2015.

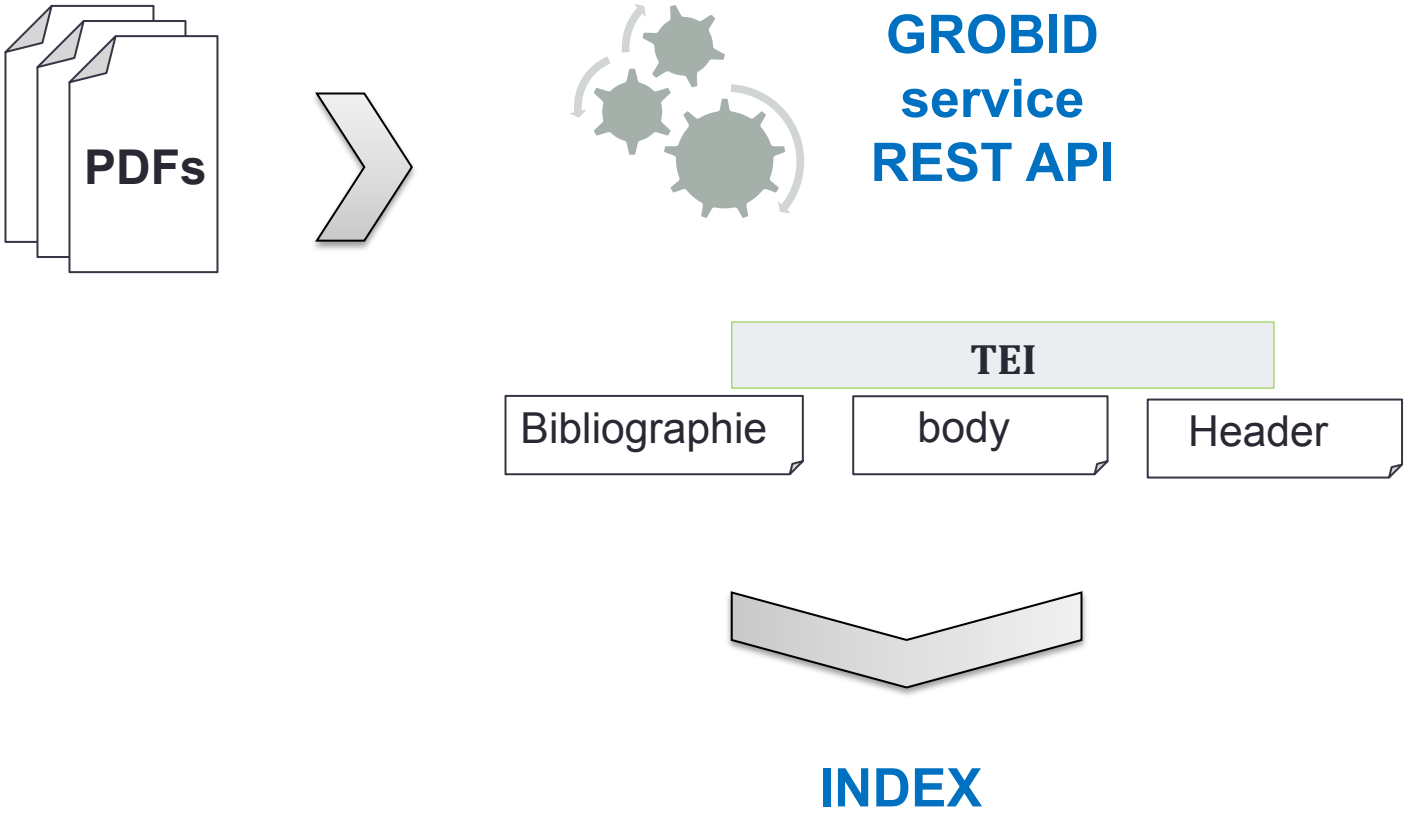
1. GROBID : GeneRation Of Bibliographic Data

Bibliographical reference parsing

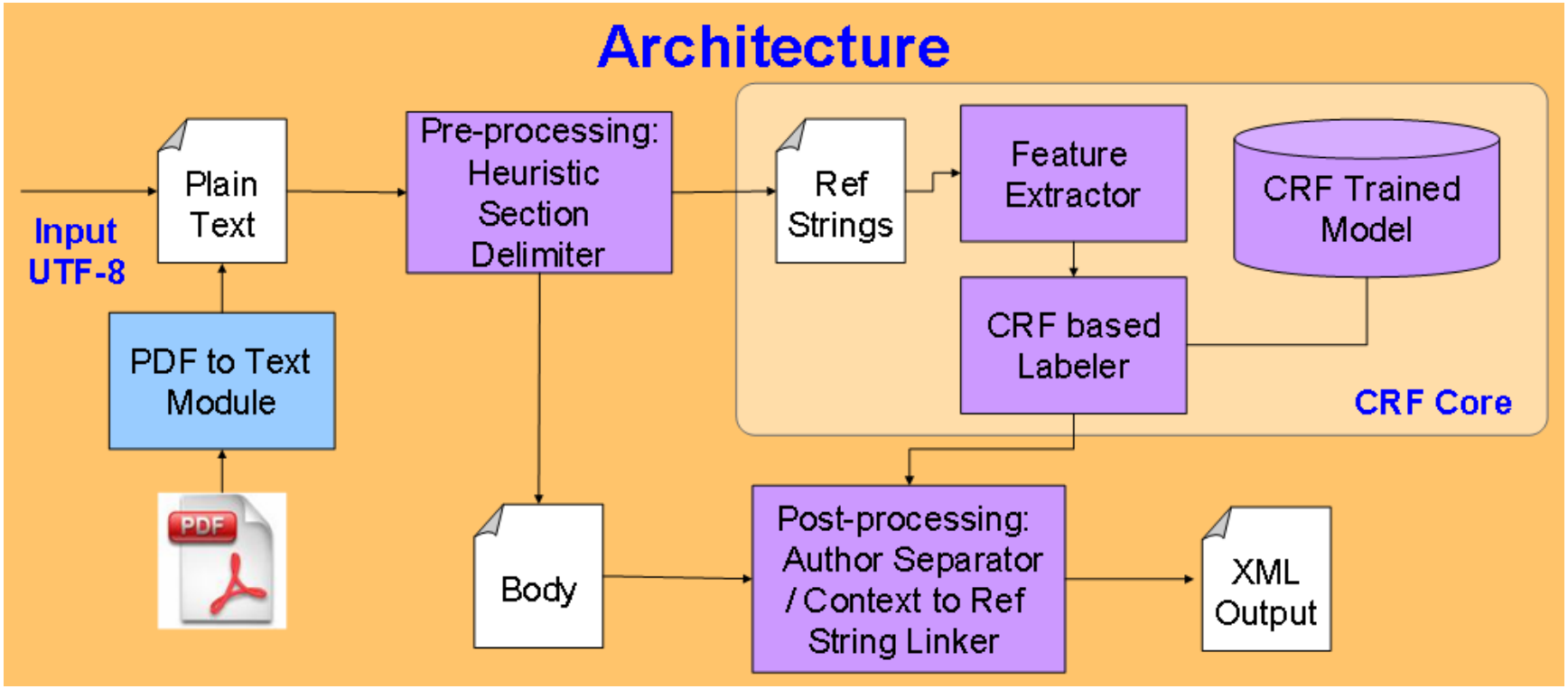


Lopez Patrice, Inria, *GROBID : from PDF to structured documents,* Avril 2015.

1. GROBID : GeneRation Of Bibliographic Data



1. Parscit



Isaac G. Councill, C. Lee Giles, Min-Yen Kan, *ParsCit: An open-source CRF reference string parsing package*, 2008

1. Extracteurs sémantiques

AlchemyAPI

```
<entity>
   <type>Person</type>
   <relevance>0.250294</relevance>
   <count>2</count>
   <text>L. A. Stoddart</text>
</entity>
<entity>
   <type>Person</type>
   <relevance>0.237201</relevance>
   <count>1</count>
   <text>R. W. Darland</text>
</entity>
```

OpenCalais

```
"http://d.opencalais.com/pershasha1/xxxxx": {
   "_typeGroup": "entities",
   "_type": "Person",
   "name": "R. L. Fowler",
   "persontype": "N/A",
   "nationality": "N/A",
   "commonname": "R. L. Fowler",
   "_typeReference": "http://
s.opencalais.com/1/type/em/e/Person",
   "instances": [
   {"detection": "[by Robertson ('39) and
Weaver, Robertson, and ]Fowler[ ('40) added
further information. Finally,
   "prefix": "by Robertson ('39) and Weaver,
Robertson, and ",
   "exact": "Fowler",
   "suffix": " ('40) added further
information. Finally, a",
   "offset": 3519,
   "length": 6 },
   … ,
   { … }
   ],
   "relevance": 0.577
}
```

2. Indexation des informations structurées

Apache Solr - Ingestion

- CSV
- Bases de données relationnelles
- File system
- Web crawl
- API / Solr XML, JSON, and javabin/SolrJ
- XML feeds (e.g. RSS/Atom),
- OAI-PMH,
- e-mail

2. Indexation des informations structurées

- **Solr – DIH *DataImportHandler***
  - *db* - database import (*example/example-DIH/hsqldb/ex.**)
  - *solr* - import from another Solr core
  - *mail* - import from IMAP
  - *tika* - import rich-content (*example/exampledocs/solr-word.pdf*)
  - *rss* - external XML feed
  - *oai* - import from OAI
  - tei – import from TEI
  - …

Solr - Flexibilité :

Schéma XML

Champs dynamiques pour l'ajout de sources hétérogènes

3. Construction de réseaux

INDEX

Données enrichies

Extraction d'entités

Référentiels d'auteurs / thématiques / organisations / Pays

Extraction de relations

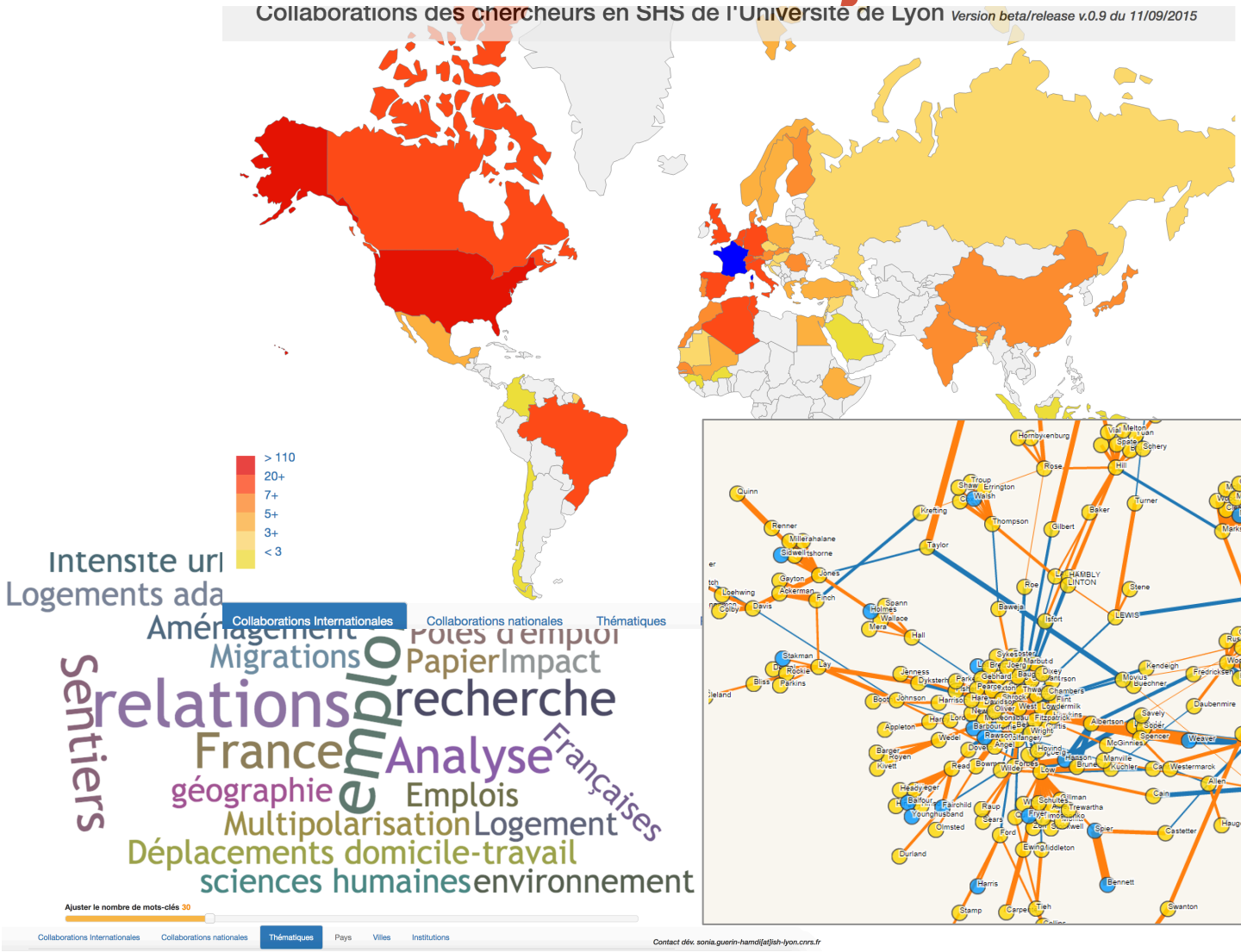
Co-écriture

Coopérations scientifiques

Citations explicites

Citations implicites

3. Visualisations – D3js



Application

Revue « Travail Humain » - JSTOR

1630 références bibliographiques + publications PDF
1933 – 2016 / 6 périodes de directions de la revue . Observation des réseaux.

p1	1933	1940
p2	1946	1966
p3	1967	1989
p4	1990	2004
p5	2005	2011
p6	2012	2016

Echantillon **2001-2015 ;** 265 publications – 3 périodes de directions.

1/ Construction du DIH + schéma XML Solr
Collecte des PDF + Indexation massive des métadonnées

2/ Prétraitement des PDFs en lot – GROBID service API REST

```
curl -v –form input=@./TH_794_0339.pdf localhost:8080/processHeaderDocument
curl -v –form input=@./TH_794_0339.pdf localhost:8080/processAffiliations
curl -v –form input=@./TH_794_0339.pdf localhost:8080/processReferences
```

3/ Préparation docker-compose.yml

```
services:
solr - image: solr:6.2.1
grobid -  image: lfoppiano/grobid:0.4.1
```

2/ Construction du graphe de collaborations. D3js

MERCI.

sonia.guerin-hamdi@ish-lyon.cnrs.fr