GROBID:
A bibliographical
& citation mining tool

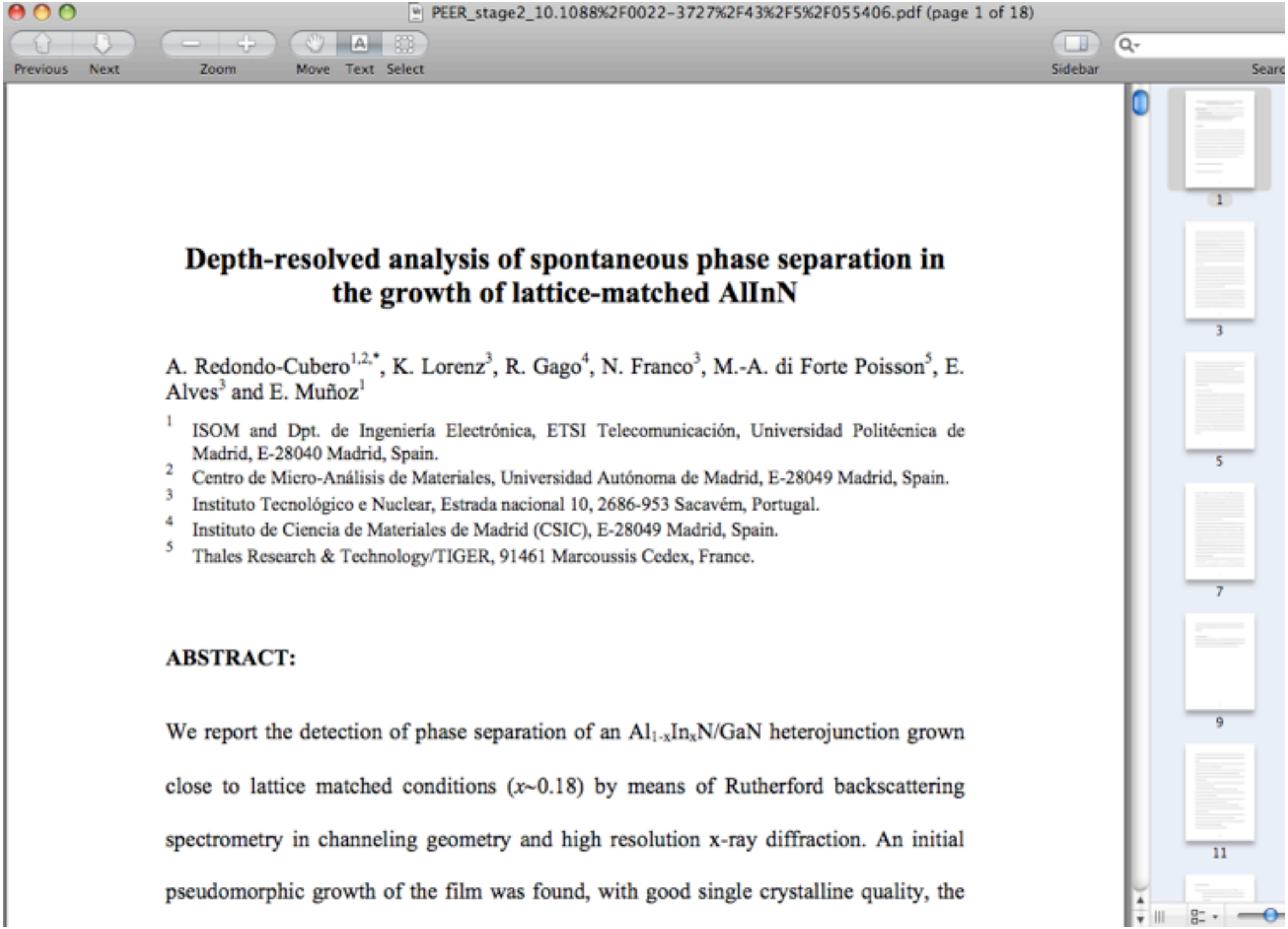
in 20 slides !

Patrice Lopez

GROBID

- **G**ene**R**ation **O**f **BI**bliographic **D**ata
- A text mining library for extracting bibliographical metadata *at large*
- Input:
  - Technical and scientific domains
  - Scholar documents, technical manuals and patents
  - Raw text or text with layout information (PDF)
- Machine learning approach: cascading of CRF models (Conditional Random Fields)
- Normalization of metadata, text and training data with the **TEI** (Text Encoding Initiative)

Example: extraction from header


PEER_stage2_10.1088%2F0022-3727%2F43%2F5%2F055406.pdf (page 1 of 18)
Previous
Next
Zoom
Move
Text
Select
Sidebar
Searc
1
3
5
7
9
11

Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

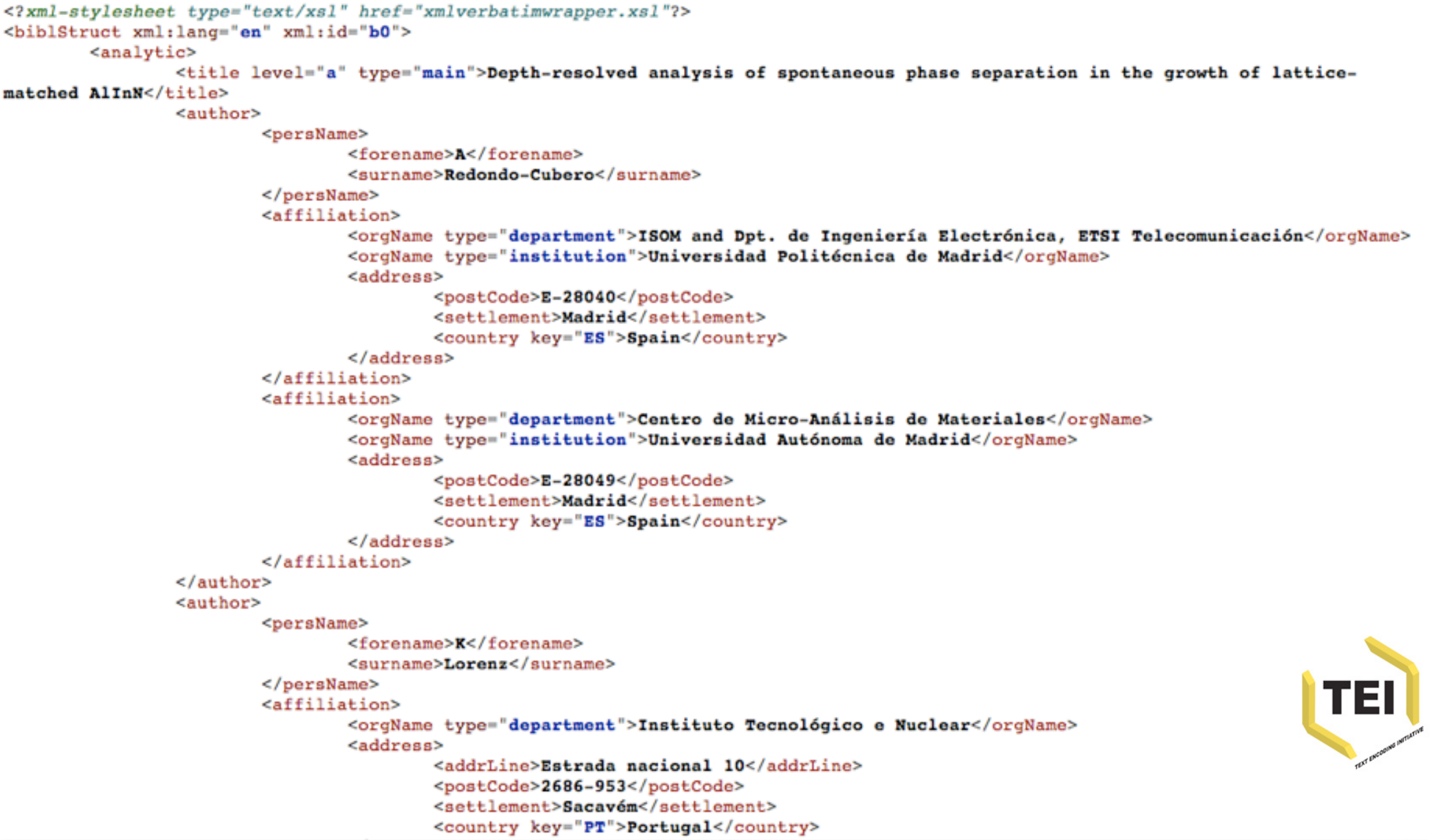
A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

[1] ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.
[2] Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.
[3] Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.
[4] Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.
[5] Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

ABSTRACT:

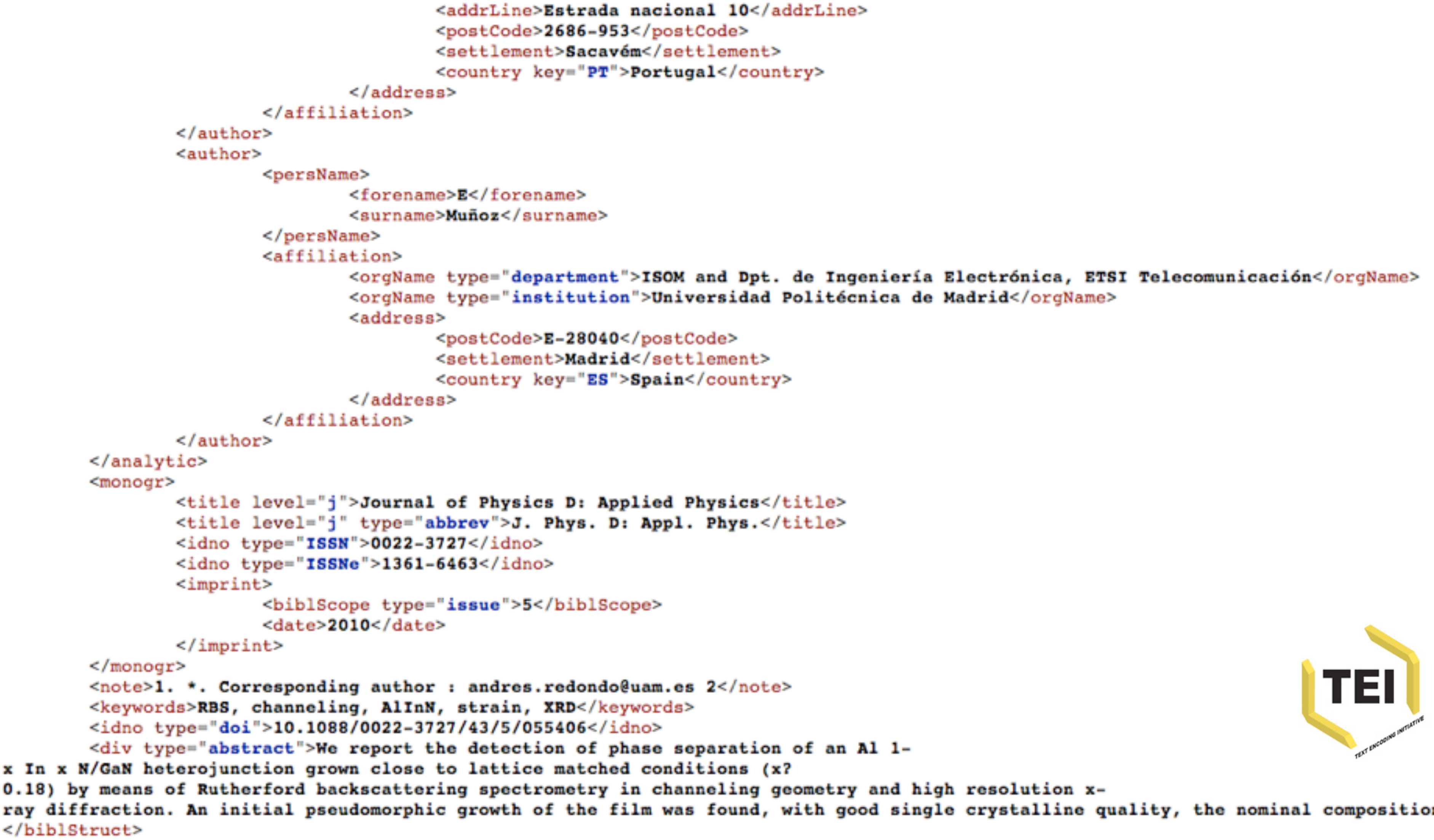
We report the detection of phase separation of an $Al_{1-x}In_xN$/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

Metadata extraction from header: TEI results



```
<?xml-stylesheet type="text/xsl" href="xmlverbatimwrapper.xsl"?>
<biblStruct xml:lang="en" xml:id="b0">
        <analytic>
                <title level="a" type="main">Depth-resolved analysis of spontaneous phase separation in the growth of lattice-
matched AlInN</title>
                <author>
                        <persName>
                                <forename>A</forename>
                                <surname>Redondo-Cubero</surname>
                        </persName>
                        <affiliation>
                                <orgName type="department">ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación</orgName>
                                <orgName type="institution">Universidad Politécnica de Madrid</orgName>
                                <address>
                                        <postCode>E-28040</postCode>
                                        <settlement>Madrid</settlement>
                                        <country key="ES">Spain</country>
                                </address>
                        </affiliation>
                        <affiliation>
                                <orgName type="department">Centro de Micro-Análisis de Materiales</orgName>
                                <orgName type="institution">Universidad Autónoma de Madrid</orgName>
                                <address>
                                        <postCode>E-28049</postCode>
                                        <settlement>Madrid</settlement>
                                        <country key="ES">Spain</country>
                                </address>
                        </affiliation>
                </author>
                <author>
                        <persName>
                                <forename>K</forename>
                                <surname>Lorenz</surname>
                        </persName>
                        <affiliation>
                                <orgName type="department">Instituto Tecnológico e Nuclear</orgName>
                                <address>
                                        <addrLine>Estrada nacional 10</addrLine>
                                        <postCode>2686-953</postCode>
                                        <settlement>Sacavém</settlement>
                                        <country key="PT">Portugal</country>
```

Metadata extraction from header: TEI results

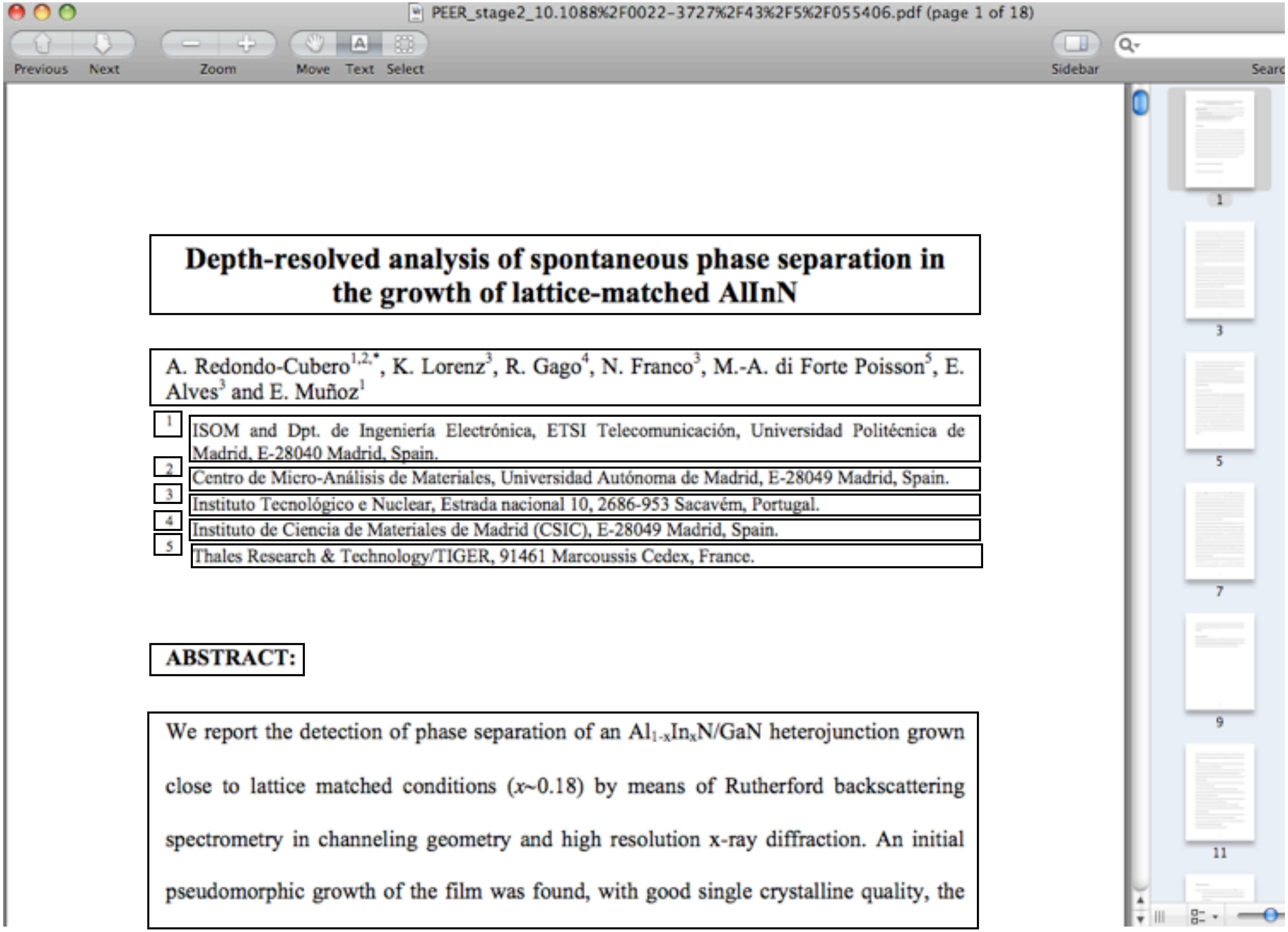


```
                    <addrLine>Estrada nacional 10</addrLine>
                    <postCode>2686-953</postCode>
                    <settlement>Sacavém</settlement>
                    <country key="PT">Portugal</country>
                </address>
            </affiliation>
        </author>
        <author>
            <persName>
                <forename>E</forename>
                <surname>Muñoz</surname>
            </persName>
            <affiliation>
                <orgName type="department">ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación</orgName>
                <orgName type="institution">Universidad Politécnica de Madrid</orgName>
                <address>
                    <postCode>E-28040</postCode>
                    <settlement>Madrid</settlement>
                    <country key="ES">Spain</country>
                </address>
            </affiliation>
        </author>
    </analytic>
    <monogr>
        <title level="j">Journal of Physics D: Applied Physics</title>
        <title level="j" type="abbrev">J. Phys. D: Appl. Phys.</title>
        <idno type="ISSN">0022-3727</idno>
        <idno type="ISSNe">1361-6463</idno>
        <imprint>
            <biblScope type="issue">5</biblScope>
            <date>2010</date>
        </imprint>
    </monogr>
    <note>1. *. Corresponding author : andres.redondo@uam.es 2</note>
    <keywords>RBS, channeling, AlInN, strain, XRD</keywords>
    <idno type="doi">10.1088/0022-3727/43/5/055406</idno>
    <div type="abstract">We report the detection of phase separation of an Al 1-
x In x N/GaN heterojunction grown close to lattice matched conditions (x?
0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-
ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the nominal compositio
</biblStruct>
```

Example: extraction from header

- Extraction of bibliographical information from the article header
- Fields: title, authors, date, abstract, location, affiliation, book title, journal title, email, publication number, web, degree, keywords, etc.
- As features, exploitation of
  - position information (begin/end of line, in the doc.)
  - lexical information (vocabulary, large gazetteers)
  - layout information (font size, font style, etc.)
- Conditional Random Fields (CRF) (Peng & McCallum 04)
- Current training corpus: 1 350 global examples + 200 affiliations/addresses blocks + 500 authors sequences, etc.

Layout & Block Analysis: XY-Cut algorithm


PEER_stage2_10.1088%2F0022-3727%2F43%2F5%2F055406.pdf (page 1 of 18)
Previous
Next
Zoom
Move
Text
Select
Sidebar
Searc
1
3
5
7
9
11

Depth-resolved analysis of spontaneous phase separation in the growth of lattice-matched AlInN

A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

1 ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, E-28040 Madrid, Spain.

2 Centro de Micro-Análisis de Materiales, Universidad Autónoma de Madrid, E-28049 Madrid, Spain.

3 Instituto Tecnológico e Nuclear, Estrada nacional 10, 2686-953 Sacavém, Portugal.

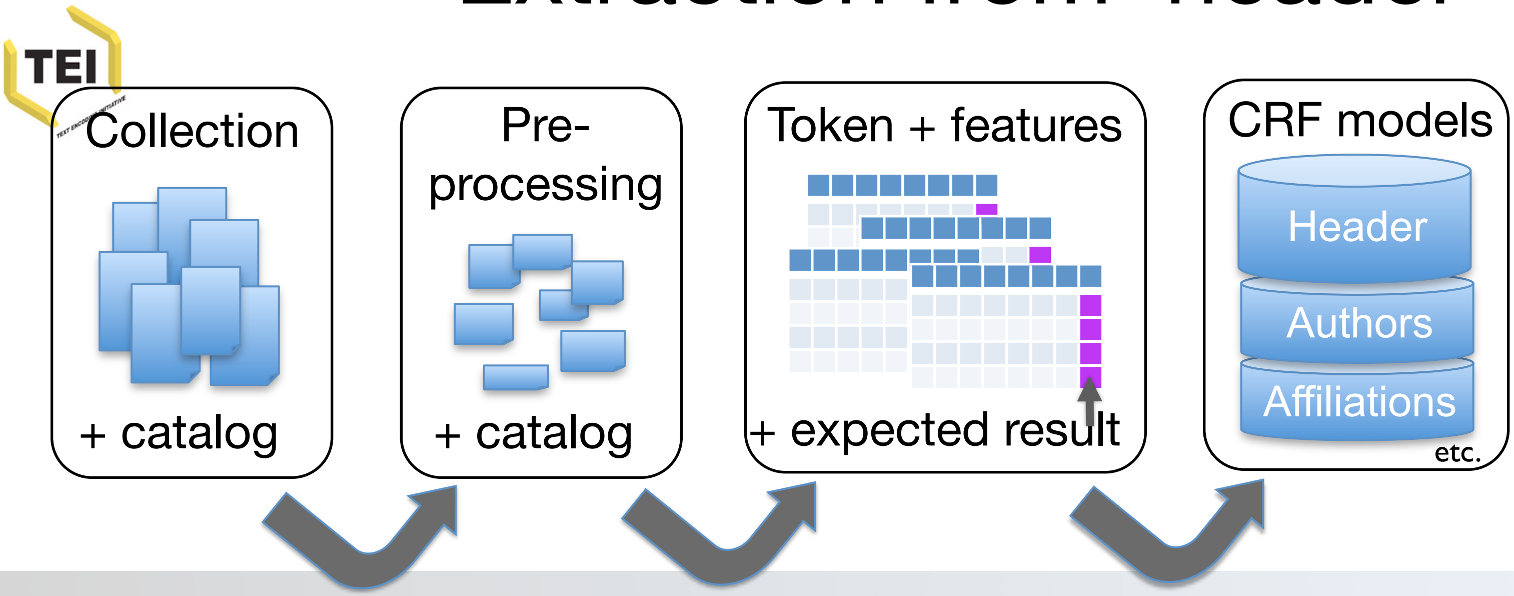
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.

5 Thales Research & Technology/TIGER, 91461 Marcoussis Cedex, France.

ABSTRACT:

We report the detection of phase separation of an Al$_{1-x}$In$_x$N/GaN heterojunction grown close to lattice matched conditions (x~0.18) by means of Rutherford backscattering spectrometry in channeling geometry and high resolution x-ray diffraction. An initial pseudomorphic growth of the film was found, with good single crystalline quality, the

Extraction from header



Collection

+ catalog

Pre-processing

+ catalog

Token + features

+ expected result

CRF models

Header

Authors

Affiliations

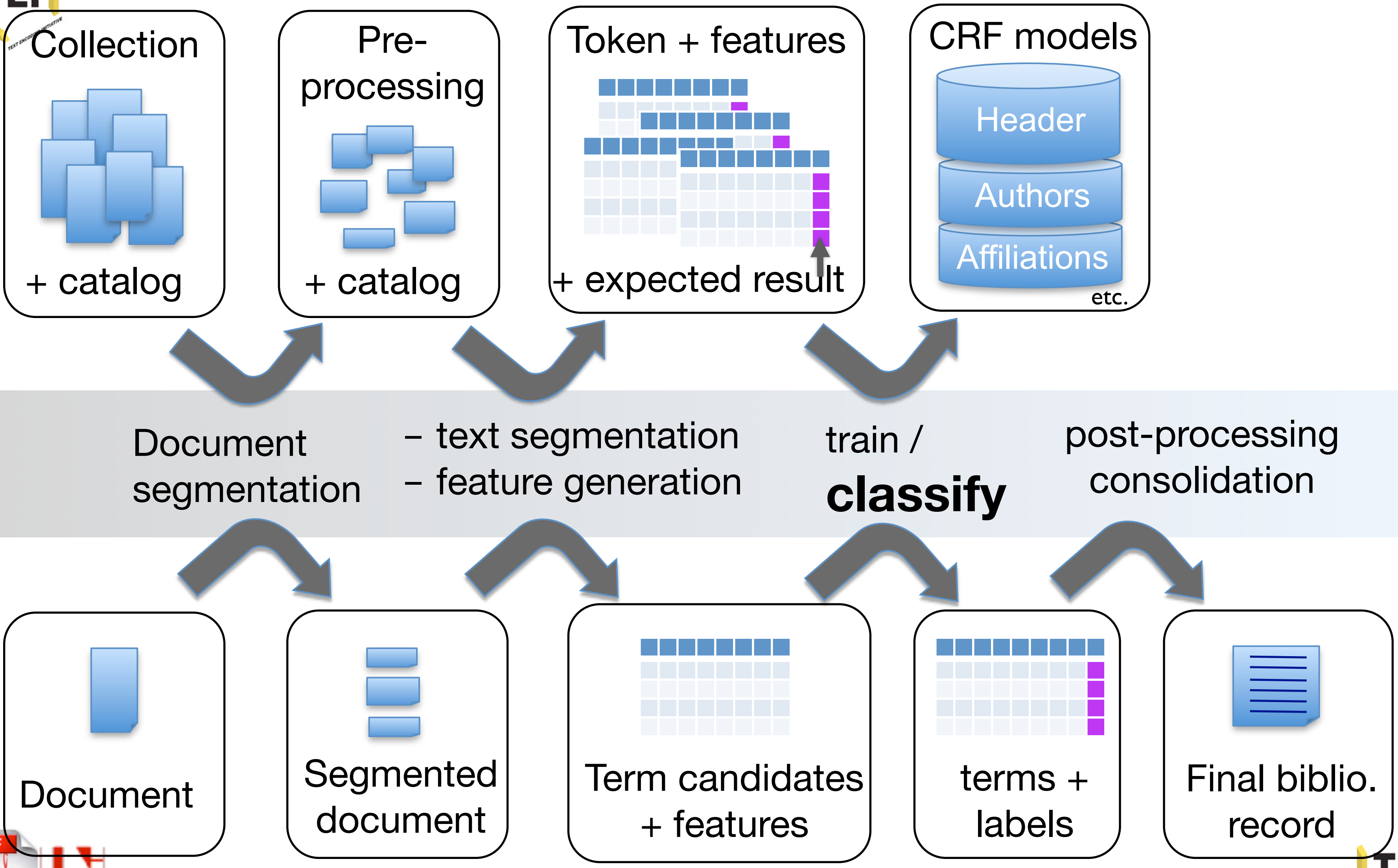
etc.

Document segmentation

- text segmentation
- feature generation

train

Extraction from header



Collection

+ catalog

Pre-processing

+ catalog

Token + features

+ expected result

CRF models

Header

Authors

Affiliations

etc.

Document segmentation

– text segmentation

– feature generation

train /

classify

post-processing consolidation

Document

Segmented document

Term candidates + features

terms + labels

Final biblio. record

Metadata consolidation

- Exploitation of external bibliographical databases for correcting/completing results based on extraction results
- **Crossref**: The full bibliographical record can be obtained based on either:
  - DOI
  - Journal title, volume, first page
  - Title + author first name → frequent!
- Other foreseen databases: xISSN, xISBN, Amazon Web Service
- Price to pay for real time processing: online requests for one consolidation between 0.8-1.5 seconds

Evaluation for headers: Corpus CORA

Features	Accuracy	Precision	Recall	F1	
Token	99.71	97.56	97.56	97.56	
Field	98.97	90.72	90.18	90.45	
Instance	–	–	–	74.91	
Instance after consolidation	–	–	–	82.20	(+9.7%)
Title	99.70	98.24	95.48	96.84	
Author	99.38	90.27	96.36	93.21	
Date	99.86	97.53	81.07	87.29	
Affiliation	99.52	98.25	93.26	95.69	
Abstract	98.95	99.64	98.81	99.22	

Grobid includes an evaluation framework for every models

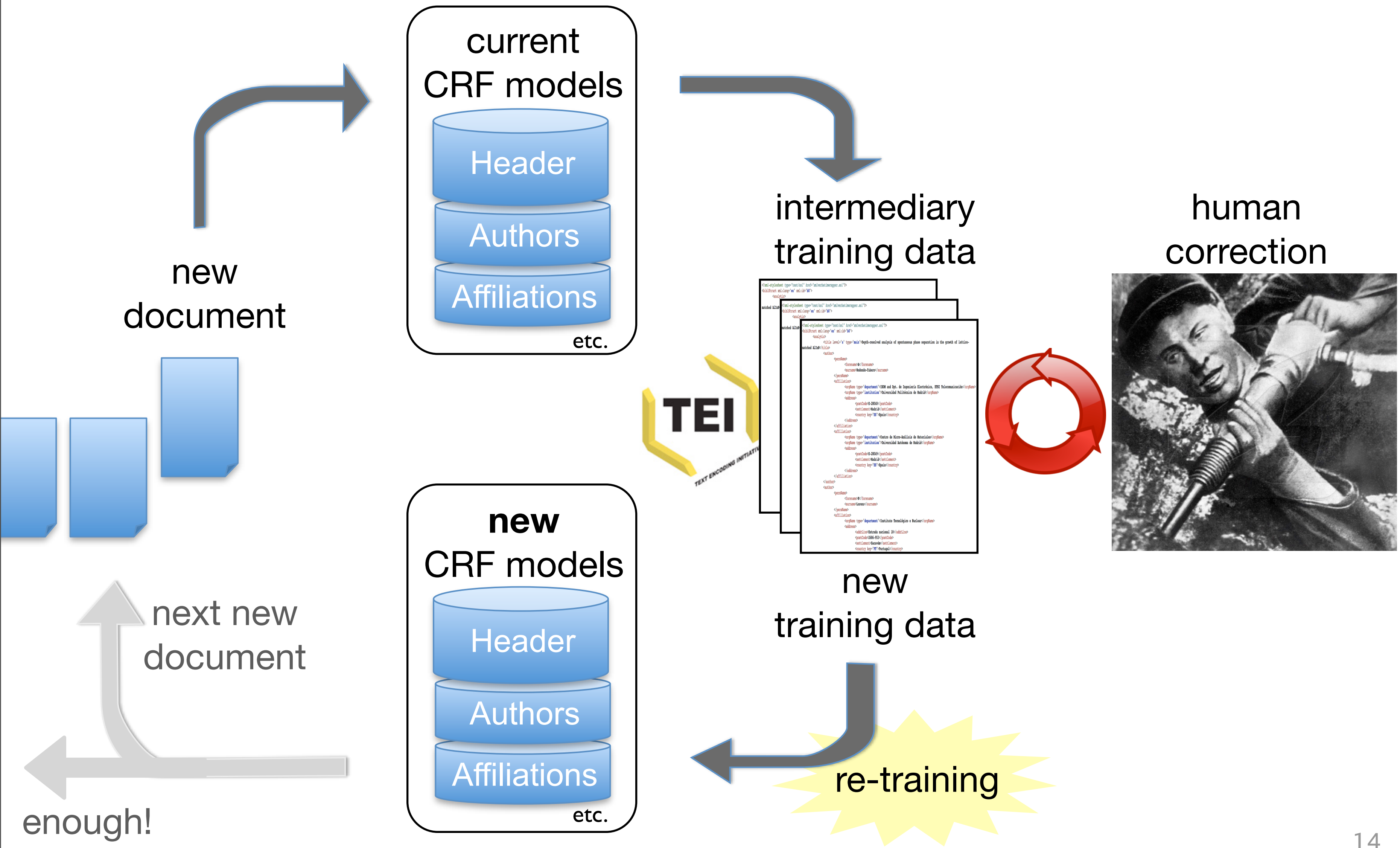
Other bibliographical extractions in GROBID

- Extraction of bibliographical references from a PDF article (with citation contexts)
- Extraction of bibliographical references in patents
  - references embedded in text
  - reference to patents
- Reference analysis: parsing of individual raw reference strings
- Key-term extraction: extraction of the most discriminant key-terms based on the specialist reader point of view

Example: bibliographical reference parsing

Raw reference	Graff, Expert. Opin. Ther. Targets (2002) 6(1): 103-113
Global citation model	CRF Citation model
First-level citation parsing	author: Graff, journal: Expert. Opin. Ther. Targets date: (2002) volume: 6(issue: 1): pages: 103-113
Second level models	CRF Author sequence CRF Date
Full parsing	author lastname: Graff journal: Expert. Opin. Ther. Targets date year: 2002 volume: 6 issue: 1 pages first: 103 pages last: 113
Consolidation	crossref DOI: 10.1517/14728222.6.1.103 Jeremy R Graff ISSN: 1472-8222 etc.
Final formatting	TEI TEXT ENCODING INITIATIVE OpenURL MPG S·F·X Get this – MIT S·F·X BIBTEX

Assisted generation of training data



new
document

current
CRF models

Header

Authors

Affiliations

etc.

intermediary
training data

```
<?xml-stylesheet type="text/xsl" href="xmlverbatimwrapper.xsl"?>
<biblStruct xml:lang="en" xml:id="b0">
	<analytic>
		<title level="a" type="main">Depth-resolved analysis of spontaneous phase separation in the growth of lattice-
matched AlInN</title>
		<author>
			<persName>
				<forename>A</forename>
				<surname>Redondo-Cubero</surname>
			</persName>
			<affiliation>
				<orgName type="department">ISOM and Dpt. de Ingeniería Electrónica, ETSI Telecomunicación</orgName>
				<orgName type="institution">Universidad Politécnica de Madrid</orgName>
				<address>
					<postCode>E-28040</postCode>
					<settlement>Madrid</settlement>
					<country key="ES">Spain</country>
				</address>
			</affiliation>
			<affiliation>
				<orgName type="department">Centro de Micro-Análisis de Materiales</orgName>
				<orgName type="institution">Universidad Autónoma de Madrid</orgName>
				<address>
					<postCode>E-28049</postCode>
					<settlement>Madrid</settlement>
					<country key="ES">Spain</country>
				</address>
			</affiliation>
		</author>
		<author>
			<persName>
				<forename>K</forename>
				<surname>Lorenz</surname>
			</persName>
			<affiliation>
				<orgName type="department">Instituto Tecnológico e Nuclear</orgName>
				<address>
					<addrLine>Estrada nacional 10</addrLine>
					<postCode>2686-953</postCode>
					<settlement>Sacavém</settlement>
					<country key="PT">Portugal</country>
```

human
correction

new
training data

re-training

new
CRF models

Header

Authors

Affiliations

etc.

next new
document

enough!

Availability

Demo !

http://scite-it.eu:8080/grobid

- Grobid is open source:
  - http://sourceforge.net/projects/grobid
- License: Apache 2.0 (do what you want...)
  - http://www.apache.org/licenses/LICENSE-2.0
- Java and C++ (CRF++) via JNI/JNATI
- xpdf is used for PDF processing
- API & RESTful interfaces (synchronous and asynchronous)
- Maven (and ant...)

- **but still work in progress...**

Authors

- The main developer is Patrice Lopez, started in 2008
- Contributors:

– Laurent Romary – Maud Medves

– Florian Zipser – Dmitry Katsubo

so with some support of ...

INRIA

Europäisches Patentamt
European Patent Office
Office européen des brevets

HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

- Grobid is used in several projects: PEER (EU), Cosmat (ANR), SLING (EU), ZNF digitalization (with the MPDL), CiteNPL (EPO)

But wait... why are you doing that?

- Cataloguing: e.g. mass digitalization
- User needs:
  - self-archiving of scholar papers by authors, e.g. in open archives
  - help when metadata are not easily available
- Extraction of additional metadata: references, keywords, etc. for enriching/correcting existing ones
  - improvement in search & retrieval
- Ease document access from citation strings (OpenURL)
- Playground for experimenting with CRF models for text mining

Ongoing & future work

- More training data and improvement of the models: the accuracy of the tool depends a lot on the volume and the diversity of the training data
- Better project packaging
- Documentation
- Full text model: full conversion of a PDF into a TEI compliant document
- Central repository of training data: sharing of training data and automatic update of CRF models

Why good bibliographical metadata are important

- Bibliographical metadata serve different purposes:
  - bibliographical item identification: this is the purpose of cataloguing
  - accessing/linking: i.e. OpenURL, exploitable by link resolvers
  - search: representation of the key information of a bibliographical item
  - interoperability: application of different services to bibliographical information

Why good bibliographical metadata are important

- Bibliographical metadata serve different purposes:
  - bibliographical item identification: this is the purpose of cataloguing - Consolidation based on external biblio DB
  - accessing/linking: i.e. OpenURL, exploitable by link resolvers - Grobid produces OpenURL results
  - search: representation of the key information of a bibliographical item - Automatic extraction of key-terms from the article content (ranked 1/19 at SemEval 2010 task 5)
  - interoperability: application of different services to bibliographical information - Grobid produces TEI and BibTex results with DOI when available via consolidation