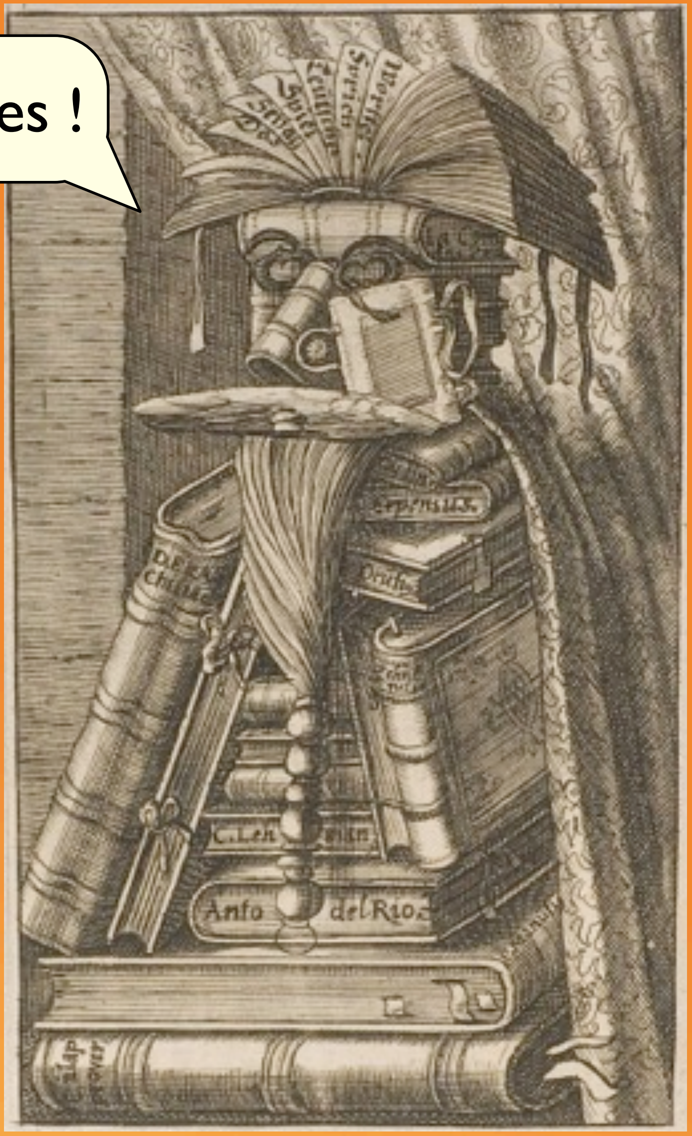
GROBID

from PDF to structured documents

in 30 slides !



inria

Patrice Lopez

April 2015

GROBID

- **G**ene**R**ation **O**f **BI**bliographic **D**ata
- A text mining library for extracting bibliographical metadata at large - started in 2008 (first as a hobby ;)
- Problem:
  - ➡ Modern digital libraries techniques require high quality metadata and full text, but we have PDF
- Goals:
  - ➡ Automatic metadata and structured content extraction from PDF
  - ➡ State-of-the-art
  - ➡ fast, robust, production-ready

GROBID

- Input:
  - Technical and scientific domains
  - Scholar documents, technical manuals and patents
  - Text with layout information (PDF) or raw text
- Machine learning approach: cascading of linear chain CRF
- Normalization of metadata
- Result and training data in TEI (Text Encoding Initiative)

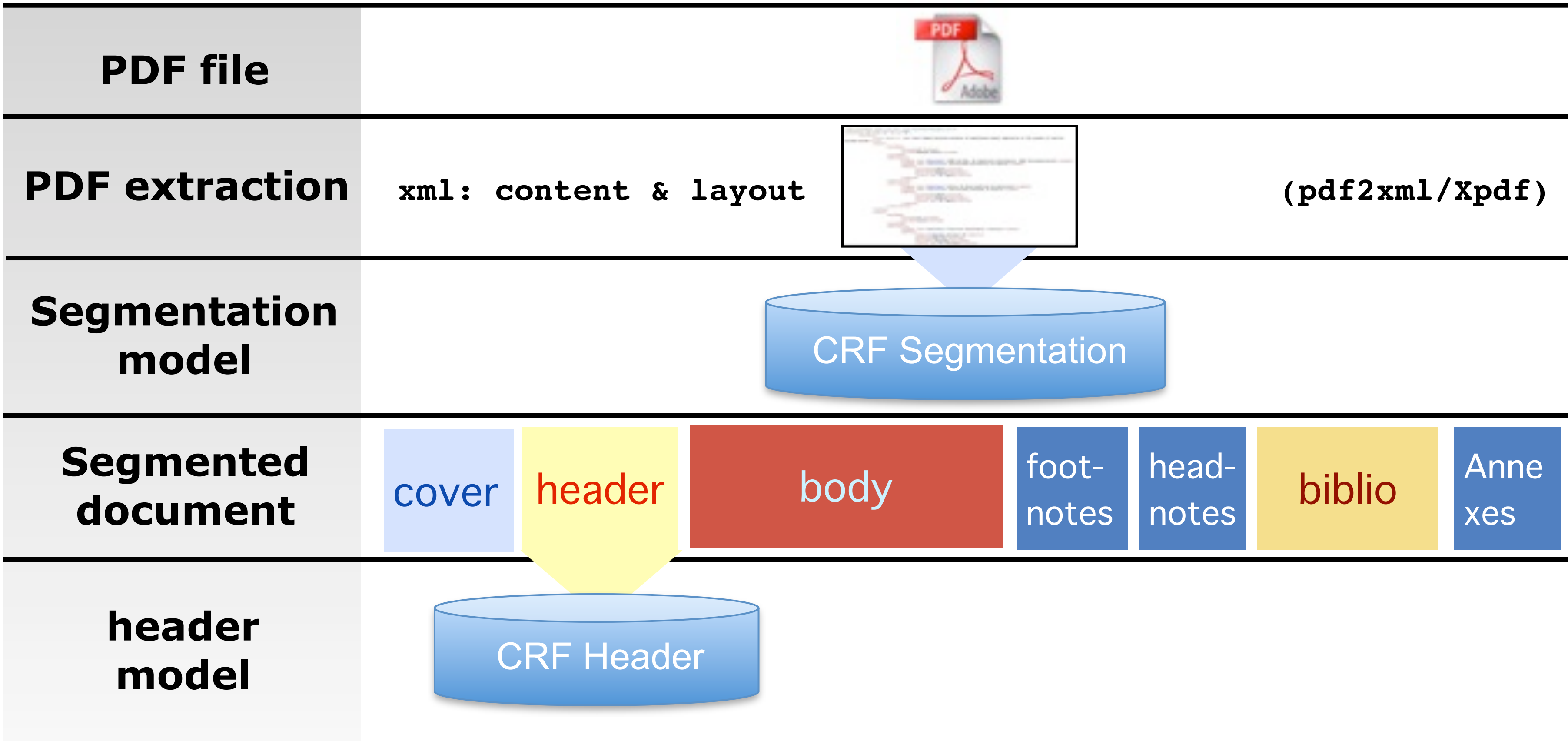
Approach

- GROBID is based on 11 different CRF models (2 for patents)
- Each model uses the same generic CRF-based framework covering training, evaluation, tokenization, decoding, etc.
- Each model has its own set of features, set of training data and normalization
- As features, exploitation of
  - ➡ position information (begin/end of line, in the doc.)
  - ➡ lexical information (vocabulary, large gazetteers)
  - ➡ layout information (font size, block, etc.)

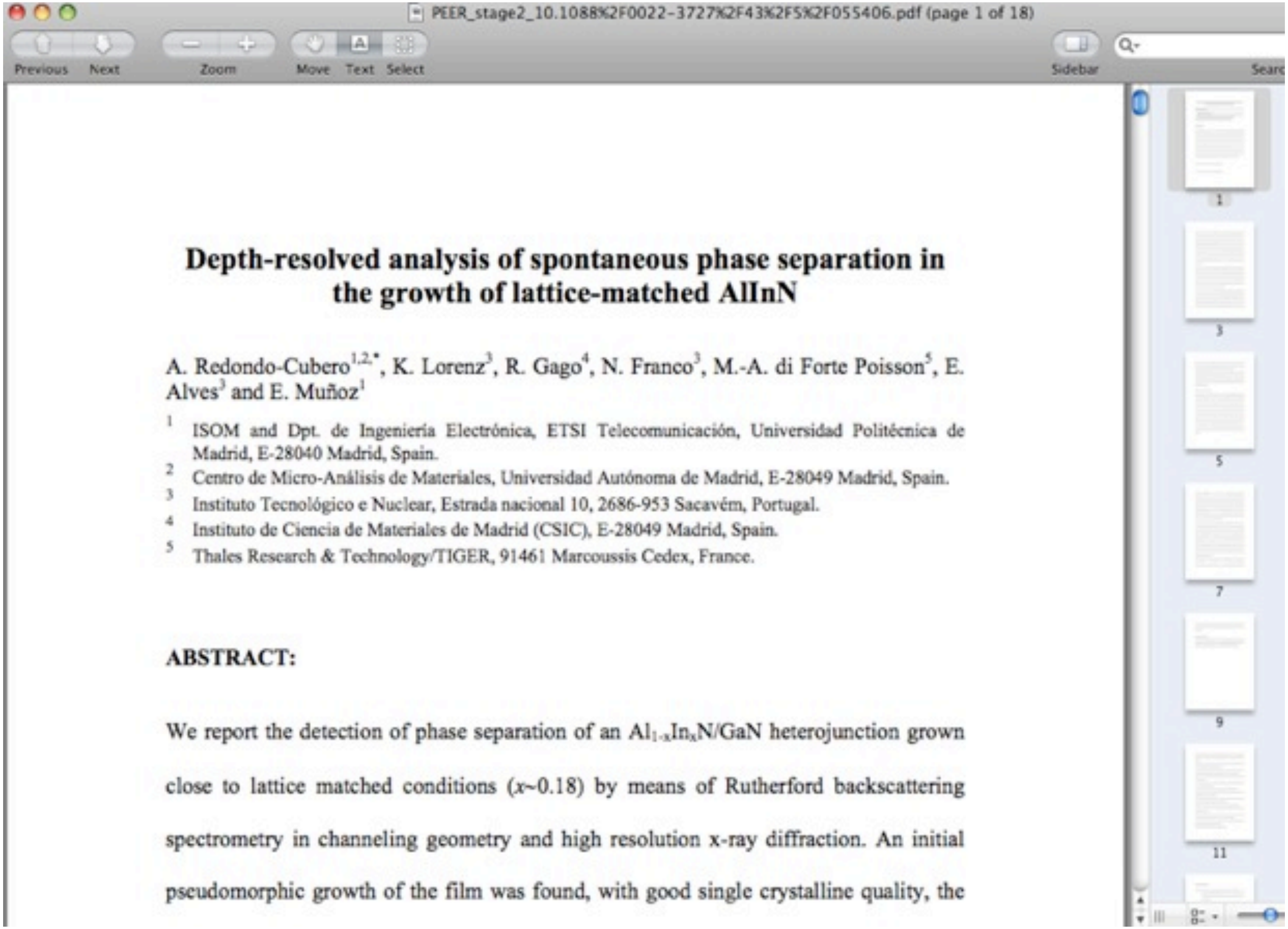
High-level segmentation (zoning)

PDF file	
PDF extraction	`xml: content & layout` `(pdf2xml/Xpdf)`
Segmentation model	CRF Segmentation
Segmented document	cover · header · body · foot-notes · head-notes · biblio · Annexes

Header processing



Example: Extraction from header



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

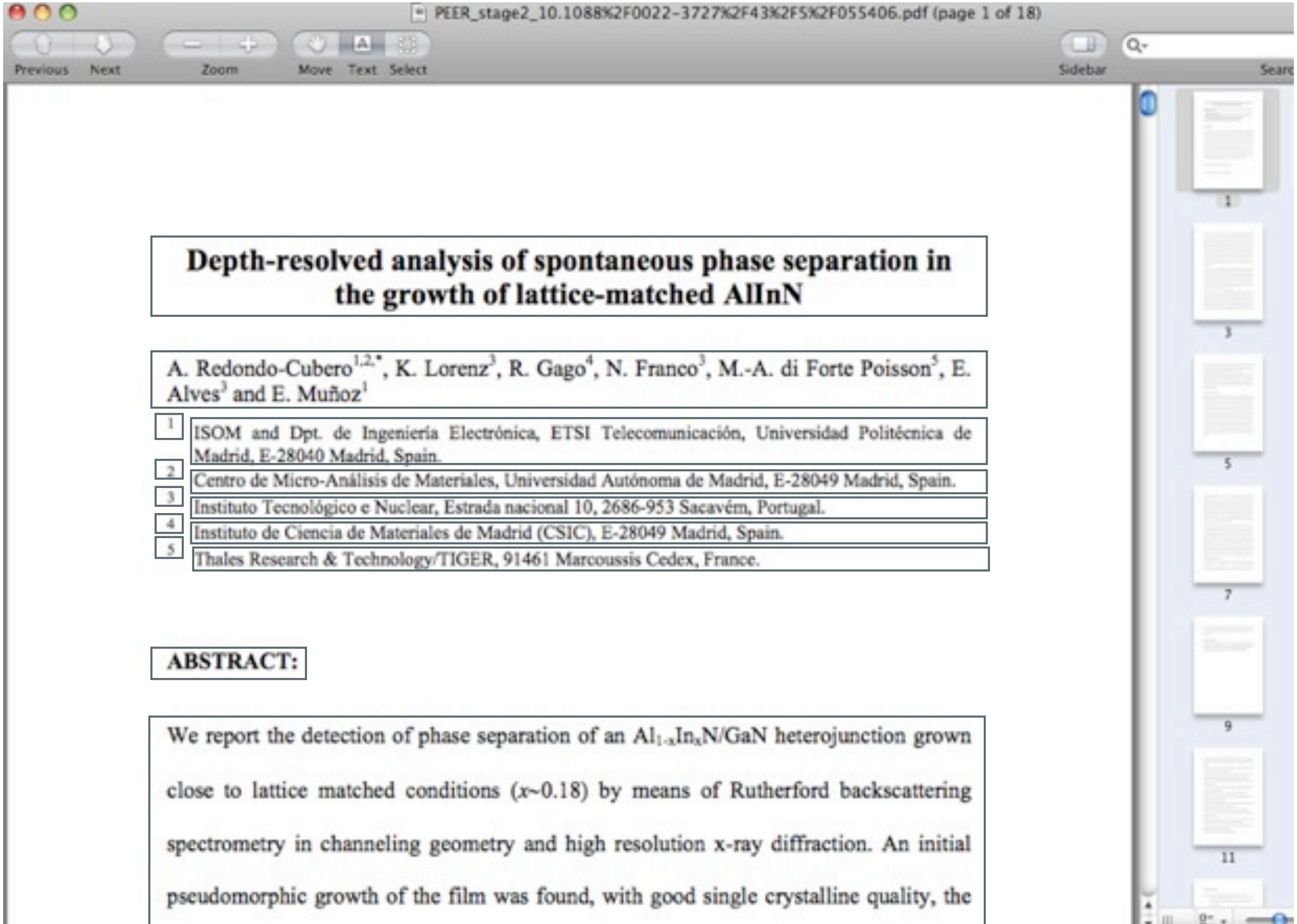
A. Redondo-Cubero[1,2,*], K. Lorenz[3], R. Gago[4], N. Franco[3], M.-A. di Forte Poisson[5], E. Alves[3] and E. Muñoz[1]

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

Example: Extraction from header



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.

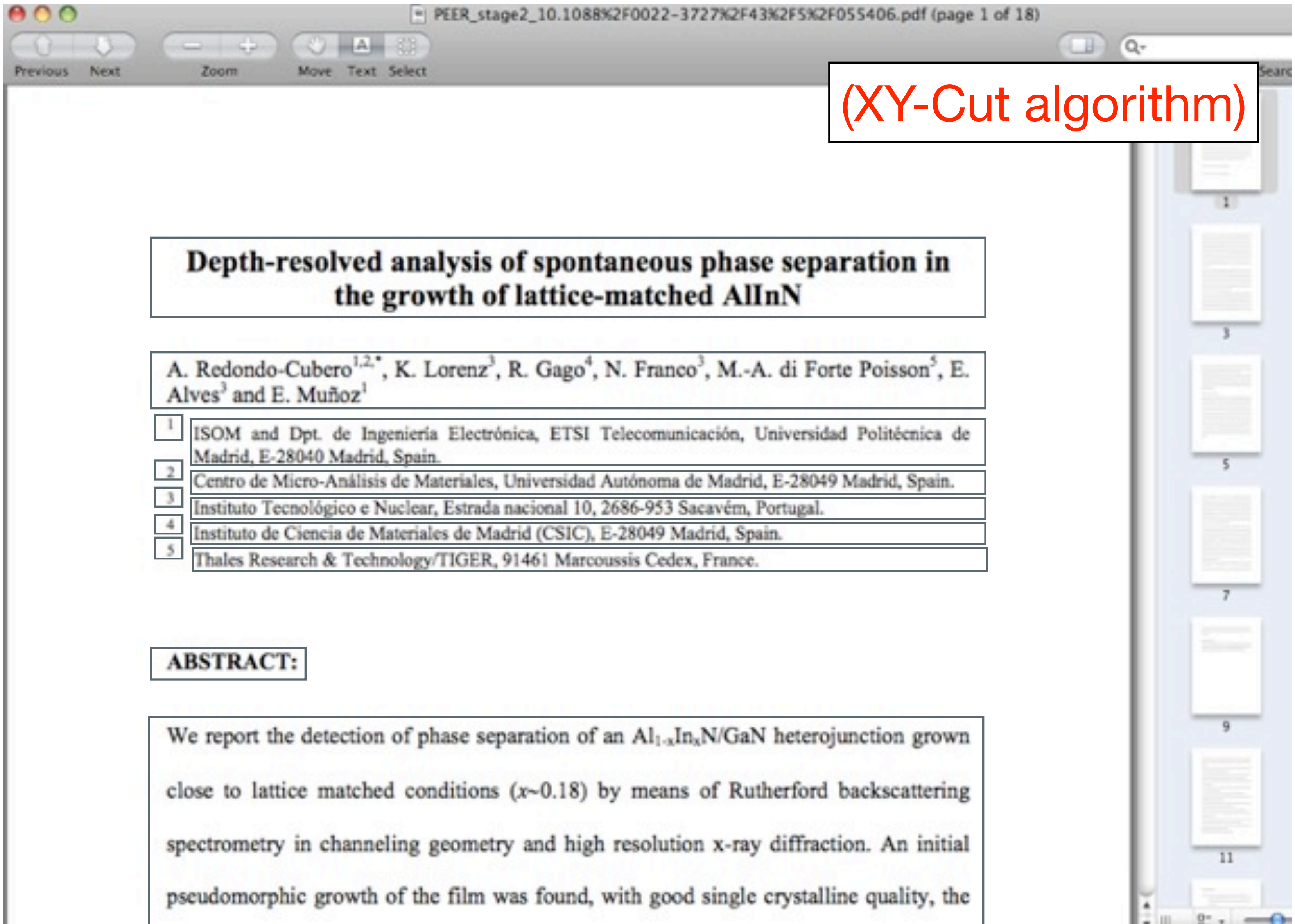
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.

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

Example: Extraction from header



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.

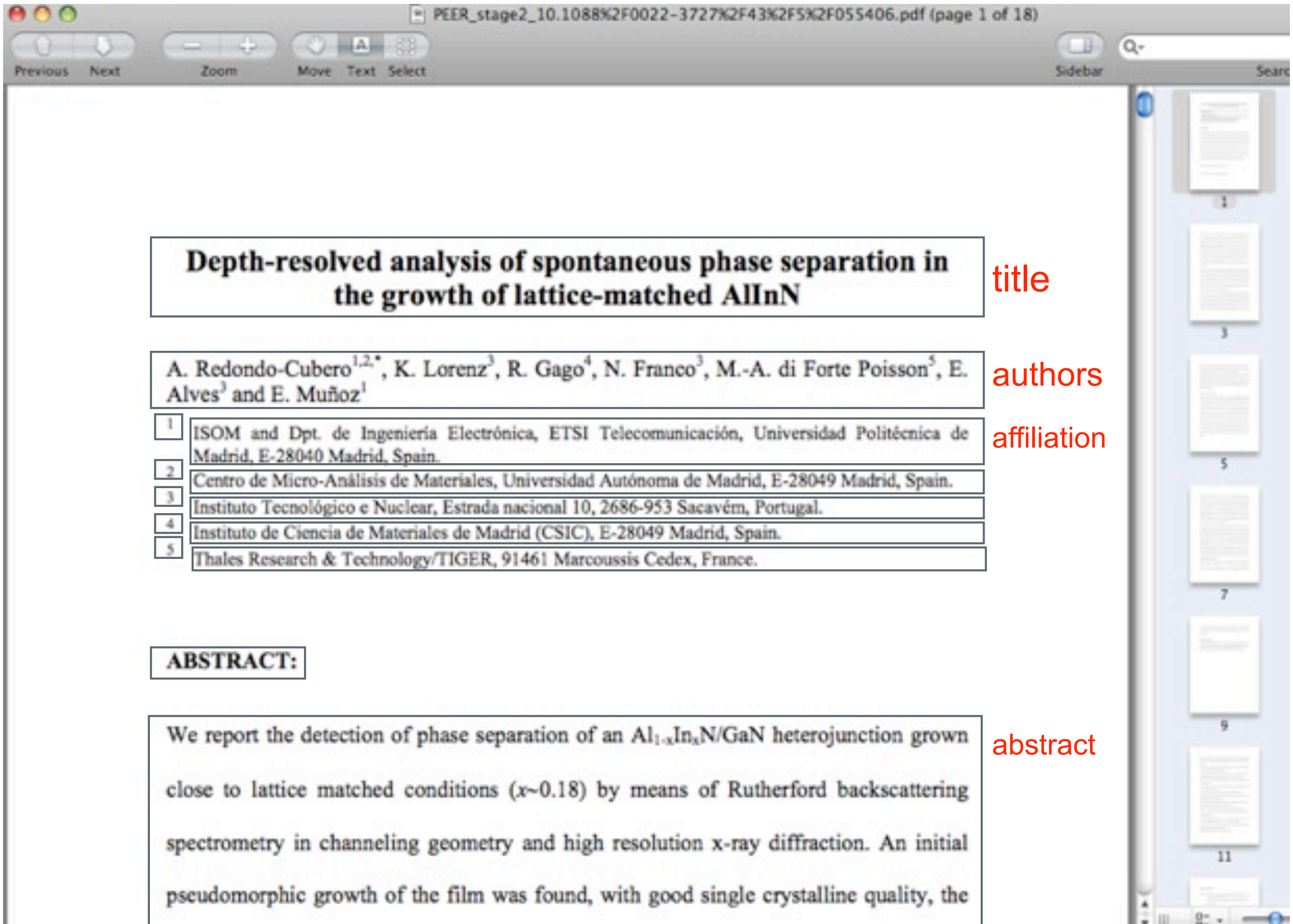
4 Instituto de Ciencia de Materiales de Madrid (CSIC), E-28049 Madrid, Spain.

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

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

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

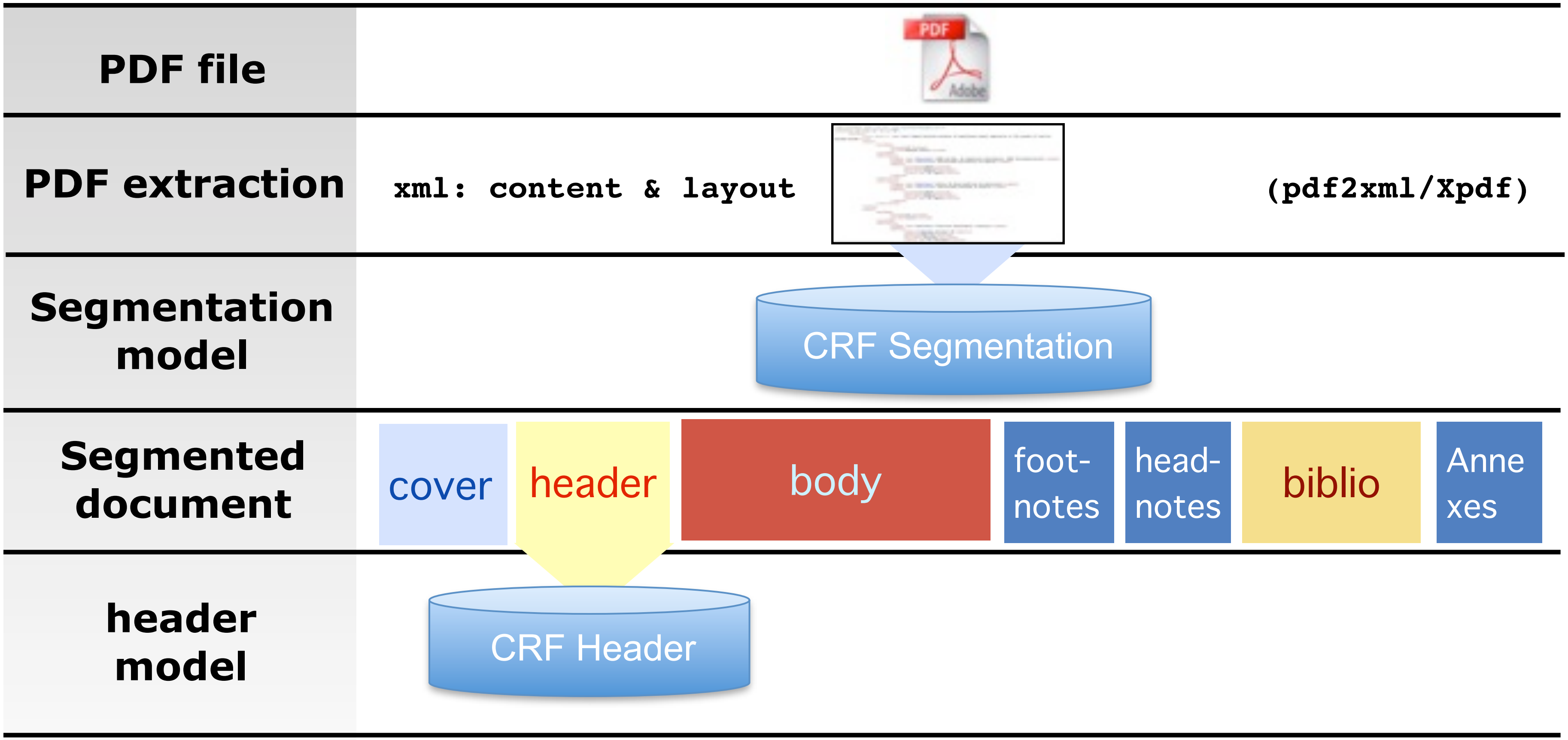
title

authors

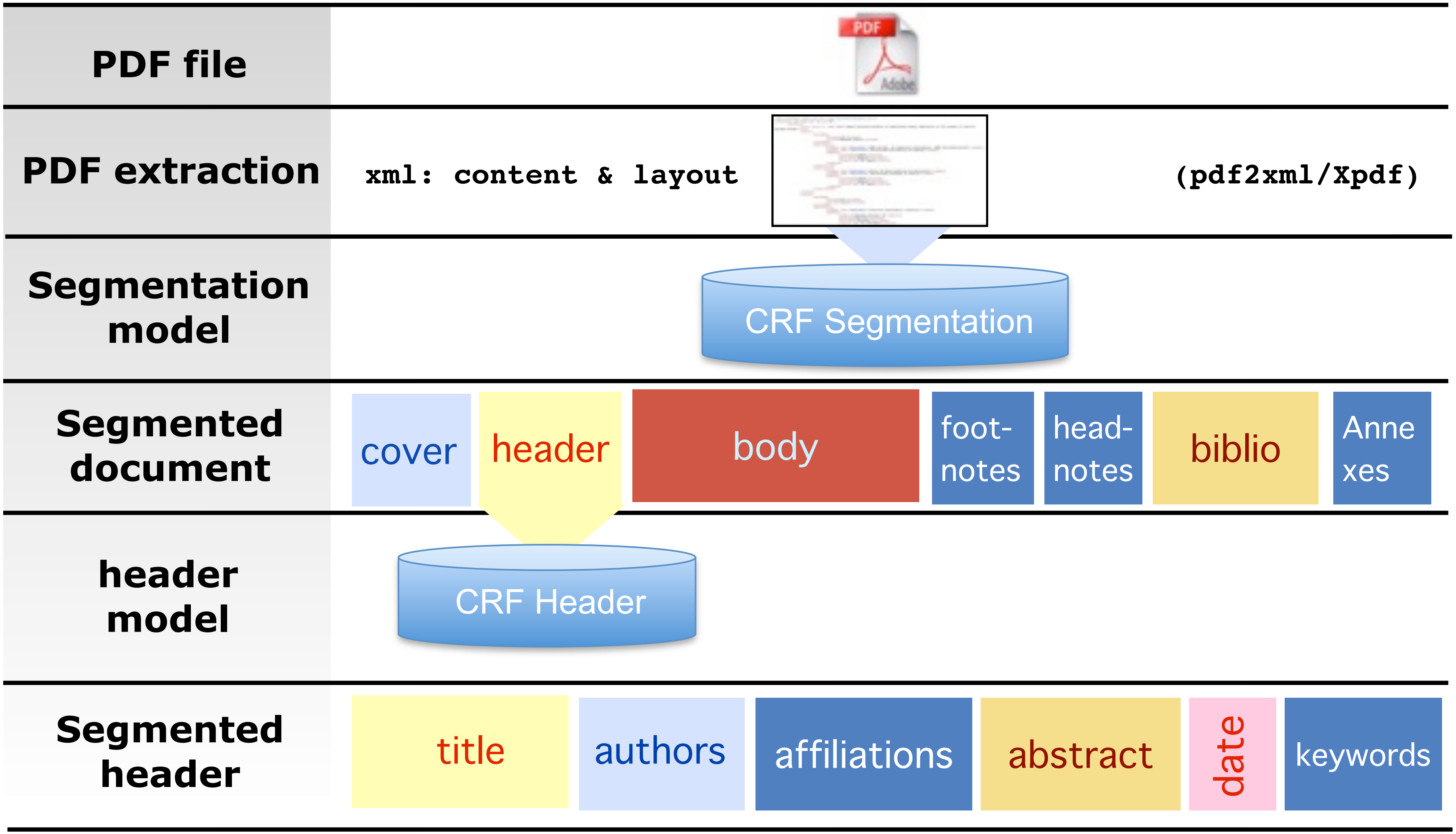
affiliation

abstract

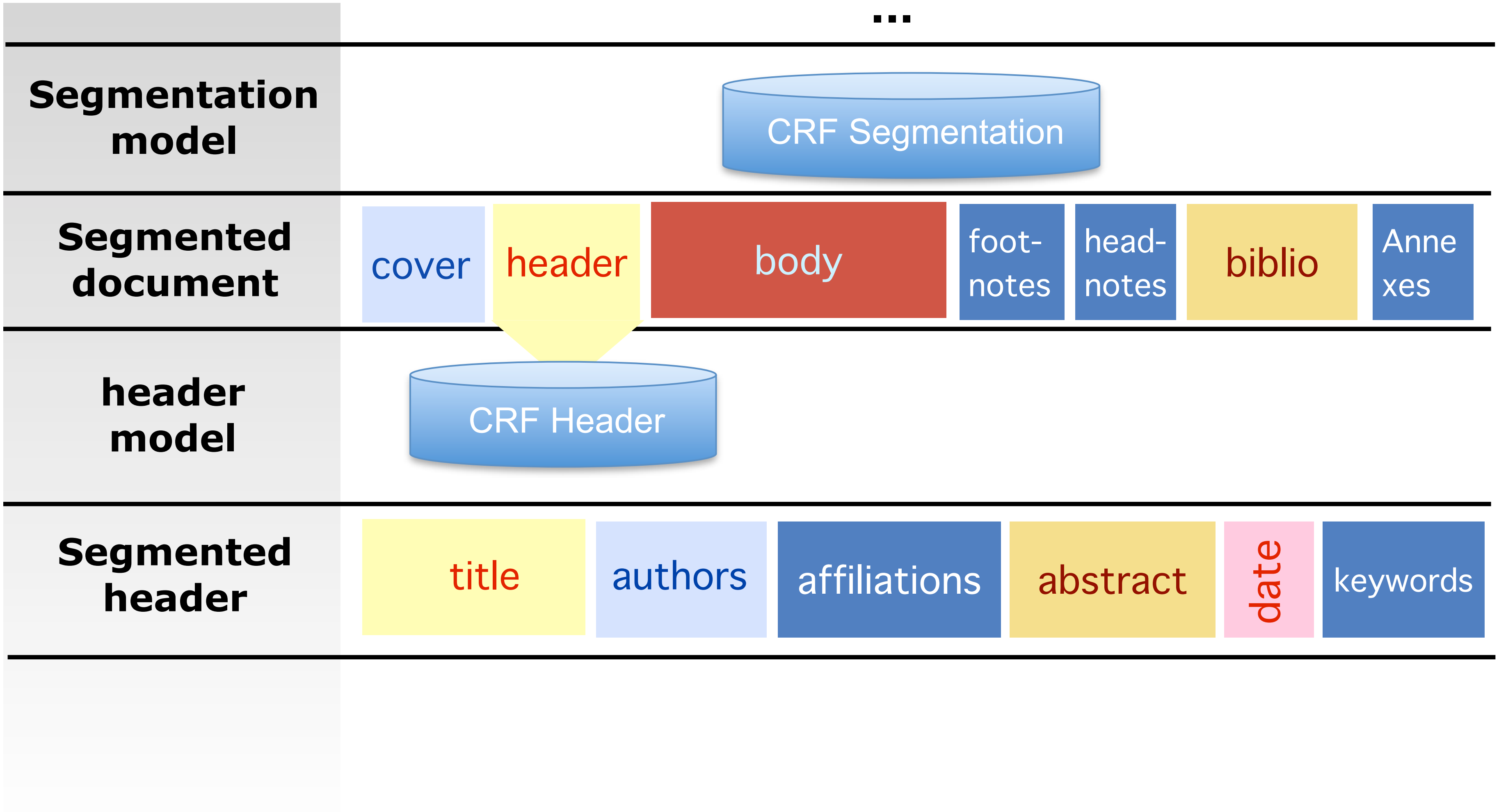
Header processing



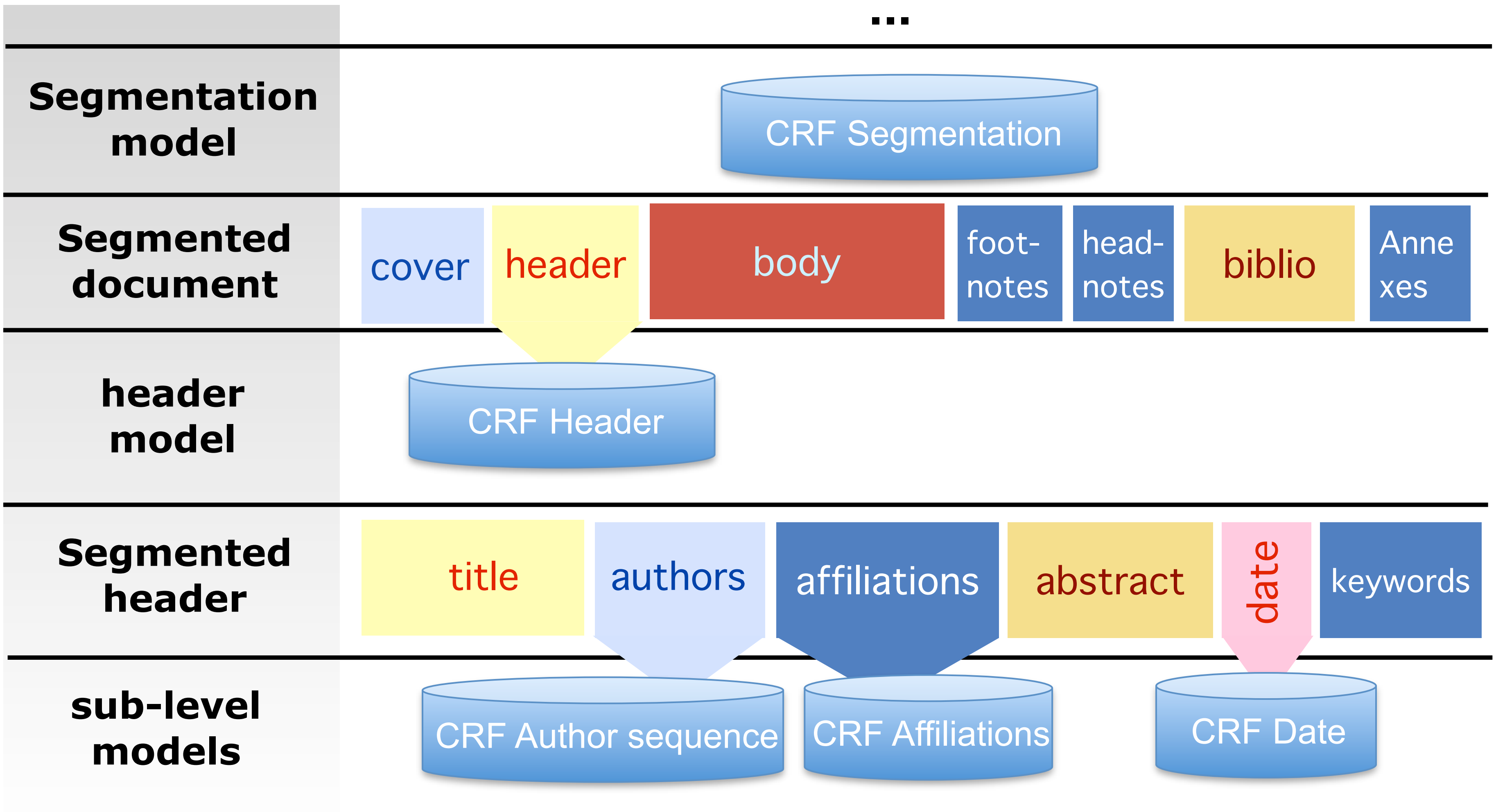
Header processing



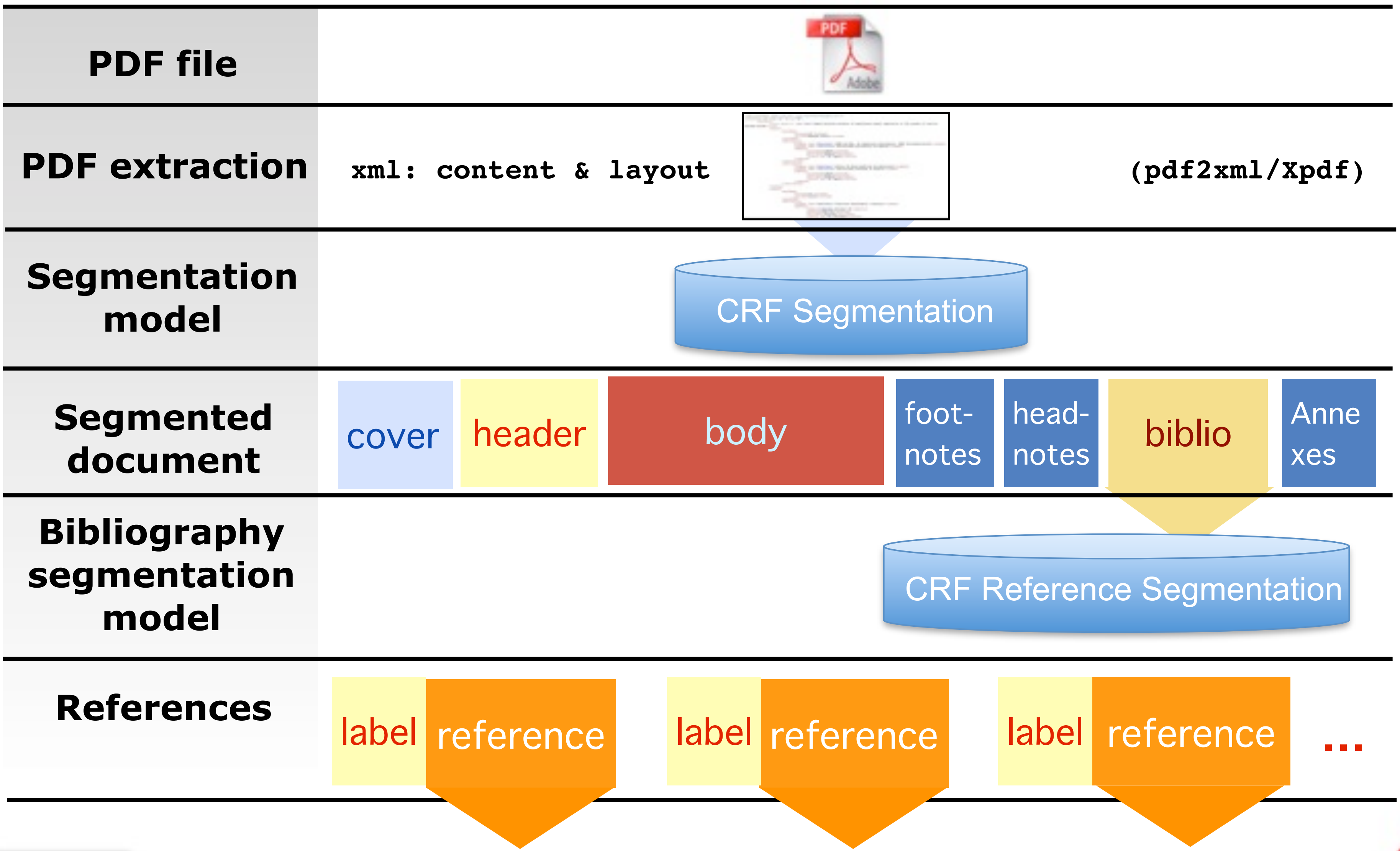
Header processing



Header processing



Bibliographical reference parsing



Bibliographical reference parsing

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 last: 113
Consolidation	crossref DOI: 10.1517/14728222.6.1.103 Jeremy R Graff ISSN: 1472-8222 etc.
Final formatting	TEI OpenURL MPG SFX BibTeX

Cascading models

- Advantages of a cascading approach:
  - ➡ hierarchical structure from “flat” linear chain CRF
  - ➡ a way to manage fine-grained structures (55 final labels, 14 intermediary labels in total in 9 models for full texts)
  - ➡ modularity: reuse of models (dates, names)
  - ➡ speed: number of labels and features for each model remains relatively low
  - ➡ training data: examples limited to one level of information

Cascading models

- Managing propagation of errors in the cascading:
  - ➡ we assume that invalid text segments for a particular level will have to be processed
  - ➡ training data in each model can include noisy input
  - ➡ spurious text segments from the upper level are "neutralized" with a dedicated label
  - ➡ still to be evaluated and tuned...

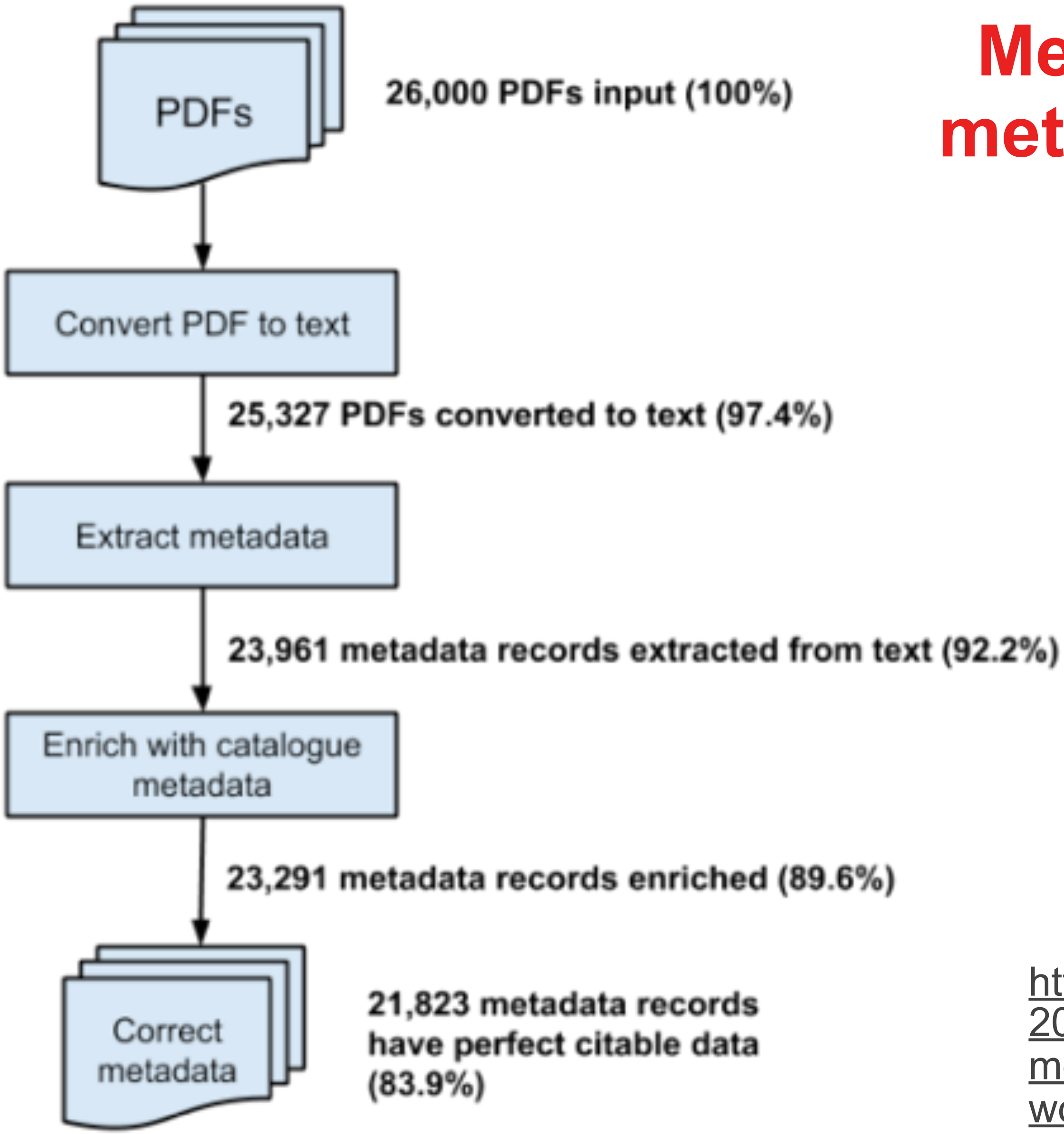
Header metadata extraction

Table 2. Results (A_{100}: First evaluation setup with 100 articles, B_{100}: Second evaluation setup with 100 articles, B_{1153}: Second evaluation setup with 1,153 articles)

	Title			Authors			Authors' last names		Abstract			Year	
	A_{100}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}	A_{100}	B_{100}	B_{1153}	B_{100}	B_{1153}
GROBID	N/A	0.92	0.92	N/A	0.83	0.83	0.90	0.91	N/A	0.75	0.74	0.64	0.69
Mendeley Desktop	N/A	0.84	0.82	N/A	0.72	0.70	0.78	0.77	N/A	N/A	N/A	0.23	0.26
ParsCit	0.59	0.52	0.54	0.47	0.29	0.31	0.36	0.37	0.49	0.31	0.26	0.06	0.07
PDFSSA4MET	0.13	0.21	0.18	0.05	0.02	0.01	0.20	0.18	N/A	N/A	N/A	N/A	N/A
PDFMeat	0.60	N/A	N/A	0.6	N/A	N/A	N/A	N/A	0.14	N/A	N/A	N/A	N/A
SciPlore Xtract	0.76	0.81	0.78	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SVMHeaderParse	0.50	0.57	0.61	0.64	0.70	0.73	0.74	0.76	0.37	0.64	0.64	0.21	0.20

from (Lipinski et al., 2013)

Mendeley's header metadata extraction evaluation



https://krisjack.wordpress.com/2015/03/12/how-well-does-mendeleys-metadata-extraction-work/

Evaluation again PubMedCentral: Header

1943 PDF from 1943 journals (2011)

Fields	Precision			Recall			f-score		
title	72.16	83.46	89.79	68.39	79.1	85.1	70.23	81.22	87.38
authors	61.41	69.27	80.24	58.26	65.72	76.13	59.79	67.45	78.13
first author	90.53	93.98	92.72	85.6	88.87	87.67	88	91.35	90.13
abstract	16.32	48.97	80.11	14.93	44.79	73.29	15.6	46.79	76.55
keywords	54.78	61.59	84.67	42.23	47.48	65.28	47.69	53.62	73.72
all fields (micro average)	59.89	72.61	85.64	54.73	66.35	78.26	57.19	69.34	81.79
all fields (macro average)	59.04	71.45	85.51	53.88	65.19	77.49	56.26	68.09	81.18

Matching
strict
soft: ignore punctuation, case and spaces
purple: Levenshtein distance ≥ 0.8

Evaluation again PubMedCentral: Header

1943 PDF from 1943 journals (2011)

Instance-level results		
Total expected instances	1943	
Total produced instances	1933	
Total correct instances	130	strict
	385	soft
	815	Levenshtein
	602	Ratcliff-Obershelp
Instance-level recall	6.73	strict
	19.92	soft
	42.16	Levenshtein
	31.14	Ratcliff-Obershelp

Matching
strict
soft: ignore punctuation, case and spaces
purple: Levenshtein distance **≥ 0.8**
grey: Ratcliff-Obershelp similarity **≥ 0.95**

Evaluation again PubMedCentral: Citations

1943 PDF from 1943 journals (2011)

Fields	Precision	Recall	f-score	
title	87.3	74.49	80.39	
authors	79.12	64.08	70.81	
first author	86.17	69.64	77.03	
date	90.66	72.87	80.79	
inTitle	81.6	73.92	77.57	
volume	91.6	76.57	83.41	
page	89.33	74.03	80.96	
all fields	86.46	72.13	78.65	micro average
	86.54	72.23	78.71	macro average

Matching
soft: ignore punctuation, case and spaces

Evaluation again PubMedCentral: Citation

1943 PDF from 1943 journals (2011)

Instance-level results		
Total expected instances	89,688	
Total produced instances	87,337	
Total correct instances	30,617	strict
	42,368	soft
	46,059	Levenshtein
	42,167	Ratcliff-Obershelp

Precision	Recall	f-score	
35.06	34.14	34.59	strict
48.51	47.24	47.87	soft
52.74	51.35	52.04	Levenshtein
48.28	47.02	47.64	Ratcliff-Obers.

Matching
strict **soft:** ignore punctuation, case and spaces **purple:** Levenshtein distance **≥ 0.8** **grey:** Ratcliff-Obershelp similarity **≥ 0.95**

Metadata consolidation

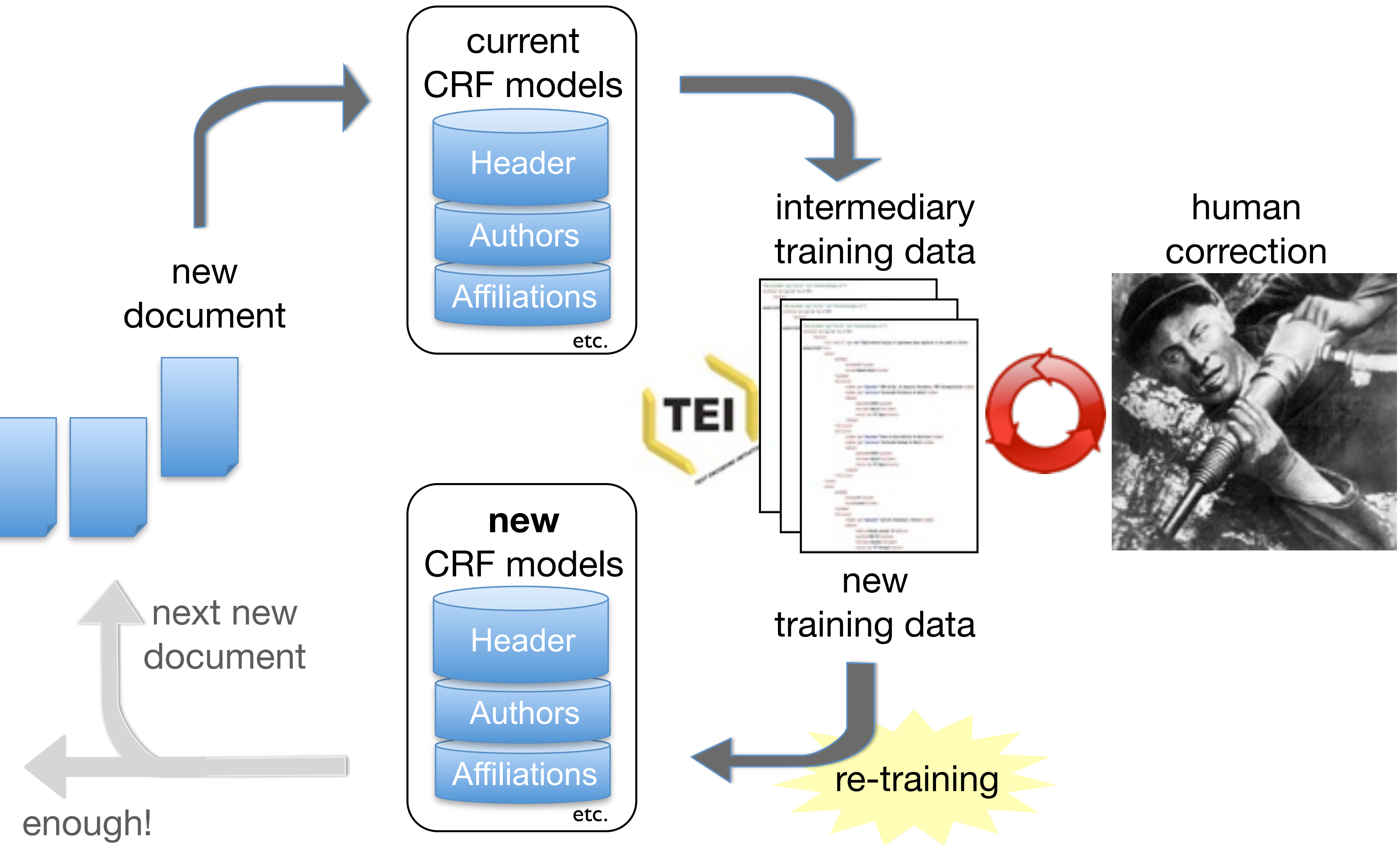
- Exploitation of external bibliographical databases for correcting/completing results based on extraction results
- Crossref: The full bibliographical record can be obtained based on either:
  - ➡ DOI
  - ➡ Journal title, volume, first page
  - ➡ Title + author first name → frequent!
- Provides ~10% improvement on header metadata extract
- Price to pay for real time processing: online requests
- Ideally use “in house” database and bibliographic deduplication techniques: ResearchGate, Mendeley, EPO
- Used at the EPO: Summon API

Training data

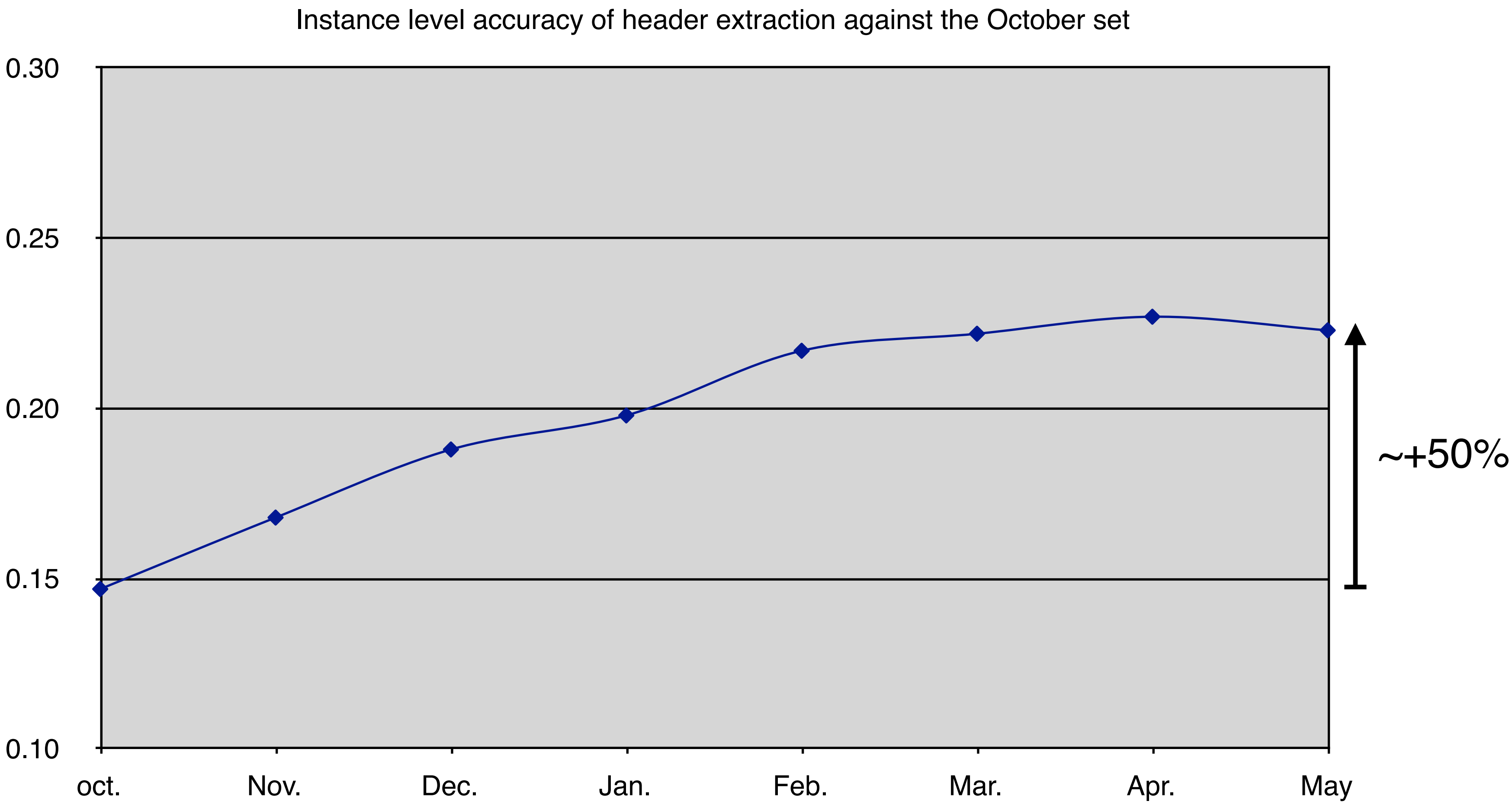
Models	# examples	exploit layout info
segmentation	121	x
header	3971	x
affiliation-address	1064	
names (header)	1297	
names (citation)	253	
date	619	
reference-segmenter	17	x
citation	4150	
fulltext (body)	8 (+13 abstracts)	x

insufficient training data (+ 2 models for patent not included here)

Assisted generation of training data



EPO project: Augmentation of training data for headers (2013-14)



Annotated headers	1530	1849	2154	2505	2855	3078	3513	3971

Technical details

- GROBID is Open Source since 02.2011

  https://github.com/kermitt2/grobid
- Apache 2.0 license
- JNI integration of the CRF libraries (CRF++, Wapiti)
- Batch, API Java & RESTful interface (with console)
- Thread-safe at parser-level
- Documentations: wiki pages, web service manual, annotation guidelines

Speed / Scaling

GROBID REST Service:

- Header: 3 PDF/s, 1 thread (MacBook)
- Citations: 12 PDF/s, 1M PDF/day on a Xeon 10 CPU E5-2660 and 10 GB memory, 3GB used in average, 9 threads (INIST)
- Full process (header, citation, fulltext): 0.6 PDF/s, 1 thread (MacBook)

Performance

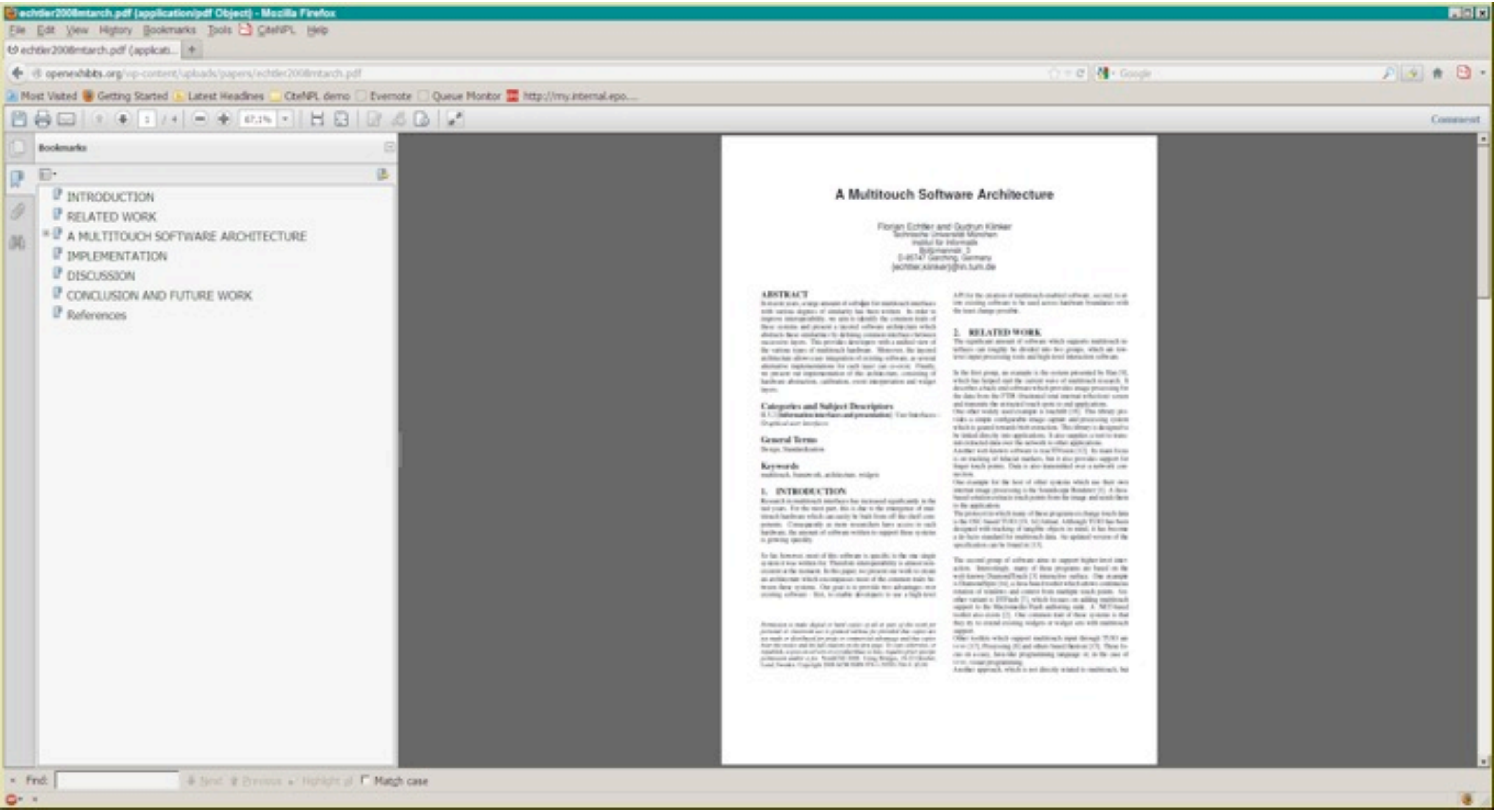
Robustness:

- for scholar literature, between 1 and 2% of the PDF parsing are failing, usually due to timeout at 20s
- an additional 1-2% of all coming PDF do not provide a usable text layer (PDF is bitmap only or the textual layer is encrypted)

Use case 1: Self-archiving of PDF

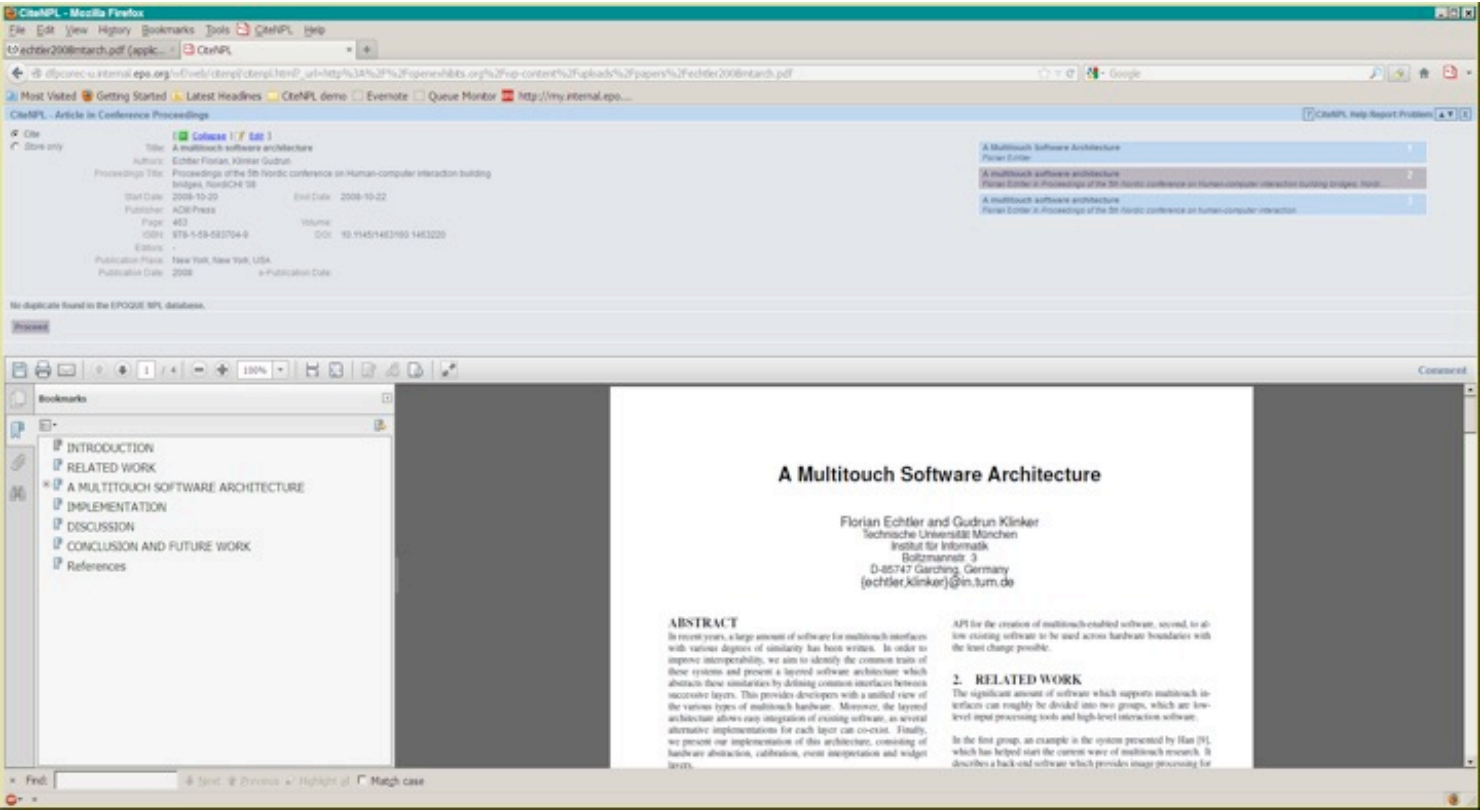
- Problem: users need to input the full bibliographical information when self-archiving or uploading a PDF
- Solution: metadata are automatically extracted from header and a pre-filled form is simply checked by the user
- This is an *online* usage of GROBID taking advantage of the sub-second PDF processing for header metadata
- In production at ResearchGate, Mendeley, HAL (French national OA archive) and EPO
- Success rate for full metadata after enrichment: 70-80%

CiteNPL - EPO



echtler2008mtarch.pdf (application/pdf Object) - Mozilla Firefox
File Edit View History Bookmarks Tools CiteNPL Help
Most Visited Getting Started Latest Headlines CiteNPL demo Evernote Queue Monitor
Bookmarks
INTRODUCTION
RELATED WORK
A MULTITOUCH SOFTWARE ARCHITECTURE
IMPLEMENTATION
DISCUSSION
CONCLUSION AND FUTURE WORK
References
A Multitouch Software Architecture
ABSTRACT
Categories and Subject Descriptors
General Terms
Keywords
1. INTRODUCTION
2. RELATED WORK
Find:
Match case

CiteNPL - EPO


CiteNPL - Mozilla Firefox
File Edit View History Bookmarks Tools CiteNPL Help
Most Visited Getting Started Latest Headlines CiteNPL demo Evernote Queue Monitor
CiteNPL - Article in Conference Proceedings
Title: A multitouch software architecture
Authors: Echtler Florian, Klinker Gudrun
Proceedings Title: Proceedings of the 5th Nordic conference on Human-computer interaction building bridges, NordiCHI '08
Start Date: 2008-10-20
End Date: 2008-10-22
Publisher: ACM Press
Page: 463
Publication Place: New York, New York, USA
Publication Date: 2008
No duplicate found in the EPOQUE NPL database.
Proceed
Bookmarks
INTRODUCTION
RELATED WORK
A MULTITOUCH SOFTWARE ARCHITECTURE
IMPLEMENTATION
DISCUSSION
CONCLUSION AND FUTURE WORK
References
A Multitouch Software Architecture
Florian Echtler and Gudrun Klinker
Technische Universität München
Institut für Informatik
Boltzmannstr. 3
D-85747 Garching, Germany
{echtler,klinker}@in.tum.de
ABSTRACT
2. RELATED WORK
Find:
Match case

CiteNPL - EPO


CiteNPL - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Most Visited Getting Started Latest Headlines
CiteNPL
CiteNPL - Article in Journal
CiteNPL Help
Title: Calibration of optically trapped nanotools [Collapse | Edit | Close CiteNPL]
Authors: Carberry D M, Simpson S H, Grieve J A, Wang Y, Schäfer H, Steinhart M, Bowman R, Gibson G M, Padgett M J, Hanna S, Miles M J
Journal: Nanotechnology
Page: 175501
Volume: 21
Issue: 17
ISSN: 0957-4484
e-ISSN: 1361-6528
DOI: 10.1088/0957-4484/21/17/175501
Publisher: IOP Pub.
Editors:
Publication Date: 30-04-2010
e-Publication Date:
Check for duplicates
Bookmarks
1. Introduction
2. Equipment and methods
3. Background theory
4. Microsphere versus tool measurements
5. Position of optical traps relative to the tool
6. Conclusions
Acknowledgments
References
IOP Publishing
Nanotechnology
Nanotechnology 21 (2010) 175501 (7pp)
doi:10.1088/0957-4484/21/17/175501
Calibration of optically trapped nanotools
D M Carberry[1], S H Simpson[1], J A Grieve[1], Y Wang[2,3], H Schäfer[2], M Steinhart[2], R Bowman[4], G M Gibson[4], M J Padgett[4], S Hanna[1] and M J Miles[1]
[1] H H Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol BS8 1TL, UK
[2] Institute for Chemistry, University of Osnabrueck, Osnabrueck, Germany
[3] State Key Laboratory of Materials-Oriented Chemical Engineering, College of Chemistry and Chemical Engineering, Nanjing University of Technology, Nanjing,
Done

Use case 2: Citation extraction at ResearchGate

- Every days, thousands of PDF are loaded either by RG users or by crawlers on OA archives
- The “acquisition” document workflow integrates Grobid for citation extraction:
  - 300K PDF are processed every months on a Hadoop cluster of 16 machines
  - Extracted citations are matched against an internal biblio. DB
- Services:
  - citation notifications for researchers
  - relevance ranking in search
- ResearchGate reported an overall Grobid failure rate of 1% on user’s self-uploaded PDF

Patrice, 3 of your publications were recently cited — Inbox

From: ResearchGate <no-reply@researchgate.net>
Subject: Patrice, 3 of your publications were recently cited
Date: November 20, 2014 12:48:19 PM GMT+01:00
To: Patrice Lopez

Hide

ResearchGate

Patrice, 3 of your publications were recently cited

NEW CITATIONS

Article: **Experiments with citation mining and key-term extraction for Prior Art Search**

Cited in 1 publication:

Conference Paper: **A standard TMF modeling for Arabic patents**
Chihebeddine Ammar, Kais Haddar, Laurent Romary

View

Article: **GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains**

On-going and future works

Ongoing projects:

- Citation extraction with INIST (France): production of training data
- CJK support, work with WIPO (Switzerland)
- Improvement of full text body restructuring

Future efforts:

- Confidence scores (with additional regression models)
- Two-stage CRF
- Document and citation classification