GROBID - Information Extraction from Scientific Publications

Laurent Romary, Patrice Lopez

GROBID - Information Extraction from Scientific Publications

by Patrice Lopez and Laurent Romary

Scientific papers potentially offer a wealth of information that allows one to put the corresponding work in context and offer a wide range of services to researchers. GROBID is a high performing software environment to extract such information as metadata, bibliographic references or entities in scientific texts.

Most modern digital library techniques rely on the availability of high quality textual documents. In practice, however, the majority of full text collections are in raw PDF or in incomplete and inconsistent semi-structured XML. To address this fundamental issue, the development of the Java library GROBID started in 2008 [1]. The tool exploits “Conditional Random Fields” (CRF), a machine-learning technique for extracting and restructuring content automatically from raw and heterogeneous sources into uniform standard TEI (Text Encoding Initiative) documents.

In the worst - but common - case, the input is a PDF document. GROBID integrates fast PDF processing techniques to extract and reorganise not only the content but also the layout and text styling information. These pieces of information are used as additional features to further improve the recognition of text structures beyond the exploitation of text only information. The tool includes a variety of CRF models specialized in different sub-structures - from high level document zoning to models for parsing dates or person names. These models can be cascaded to cover a complete document.

The first and most advanced model is dedicated to the header of a scientific or technical article and is able to reliably extract different metadata information such as titles, authors, affiliations, address, abstract, keywords, etc. This information is necessary in order to identify the document, make it citable, and use it in library systems. Following an evaluation carried out for this task in 2013 by [2], GROBID provided the best results over seven existing systems, with several metadata recognized with over 90% precision and recall. For header extraction and analysis, the tool is currently deployed in the production environments of various organizations and companies, such as the EPO, ResearchGate, Mendeley and finally as a pre-processor for the French national publication repository HAL.

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

title

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]

authors

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.

affiliation

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

abstract

Figure 1: Block segmentation of PDF documents before construing content.

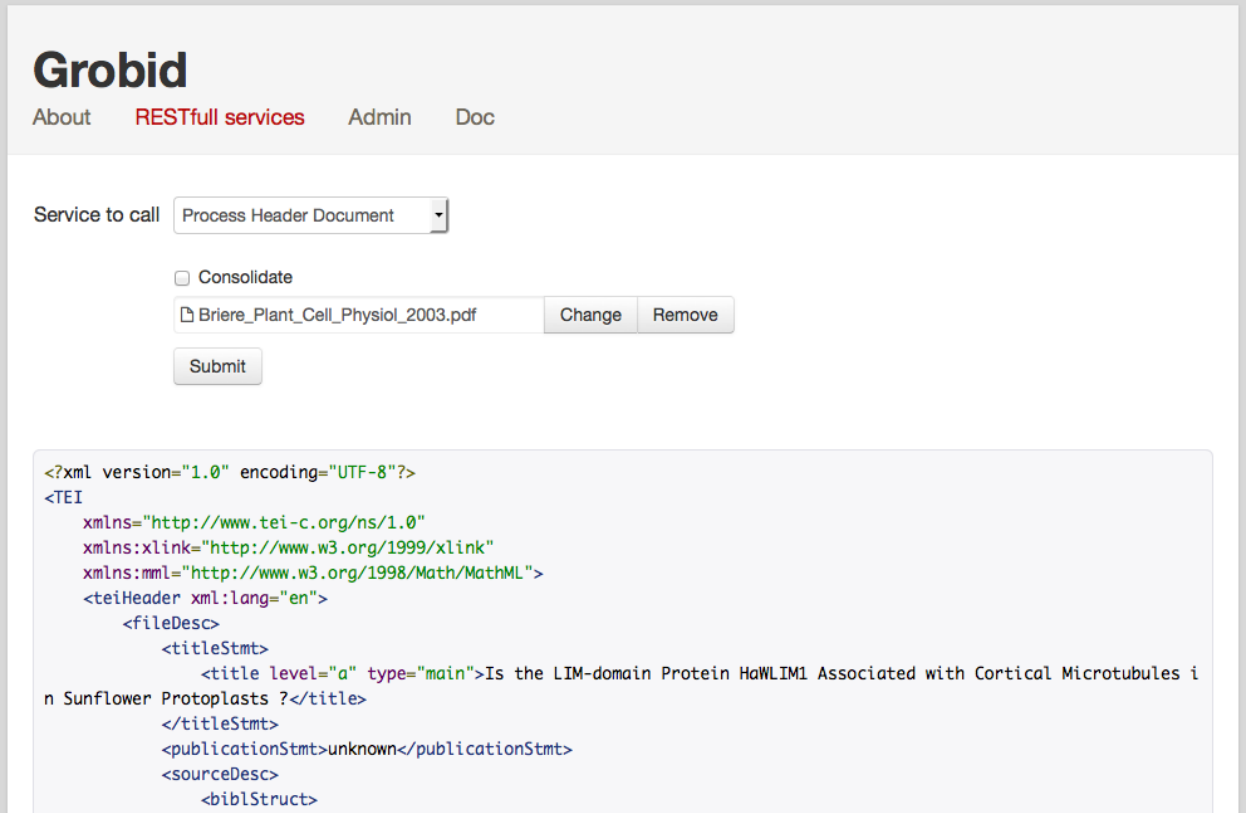

Figure 2: Online service for GROBID with TEI compliant export.

GROBID also includes a state of the art model for the extraction and the recognition of bibliographic citations. The references present in an article or patent are identified, parsed, normalized and can be matched with a standard reference database such as CrossRef or DocDB (patents). Citation information is considered very useful for improving search ranking and makes it possible to run bibliographic studies and graph-based social analyses. For instance, the citation notification service of ResearchGate uses GROBID bibliographic reference extraction to process every uploaded article. When an existing publication of a registered member is identified, the member can be informed where and how his work has been cited.

More challenging, the restructuring of the body of a document (potentially including figures, formula, tables, footnotes, etc.) is continually improving and is currently the object of the semi-automatic generation of more training data. Although more experimental, it can provide to a search engine for scientific literature richer and better text content and structures than basic PDF extractors (e.g., pdftotext, Apache TIKA or PDFBox).

The objectives of GROBID are still mainly research challenges, but significant efforts have also been dedicated to engineering. The tool can be used as web services or batch and is fast enough to scale to millions of documents in reasonable time and cluster. On a single low end hardware, GROBID processes, on average, three PDF documents per second or 3000 references in less than 10 seconds. Since 2011, the tool has been available as Open Source (Apache 2 licence) to any developers/third parties (see link below). New contributors are of

course welcome. Version 0.3 of the tool has just been released, and its development will continue over the next few years with the participation of various national and international collaborators.

Links:
Text Encoding Initiative: http://www.tei-c.org
https://github.com/kermitt2/grobid

References:
[1] P. Lopez: “GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications”, in proc. of ECDL 2009, 13th European Conference on Digital Library, Corfu, Greece, 2009.

[2] M. Lipinski, et al.: “Evaluation of header metadata extraction approaches and tools for scientific PDF documents”, in proc. of the 13th ACM/IEEE-CS joint conference on Digital libraries (JCDL '13), ACM, New York, NY, USA, 385-386, 2013. DOI=10.1145/2467696.2467753, http://doi.acm.org/10.1145/2467696.2467753.

[2] P. Lopez, L. Romary: “HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID”, SemEval 2010 Workshop, Uppsala, Sweden. https://hal.inria.fr/inria-00493437

Please contact:
Laurent Romary
Inria, France
E-mail: laurent.romary@inria.fr