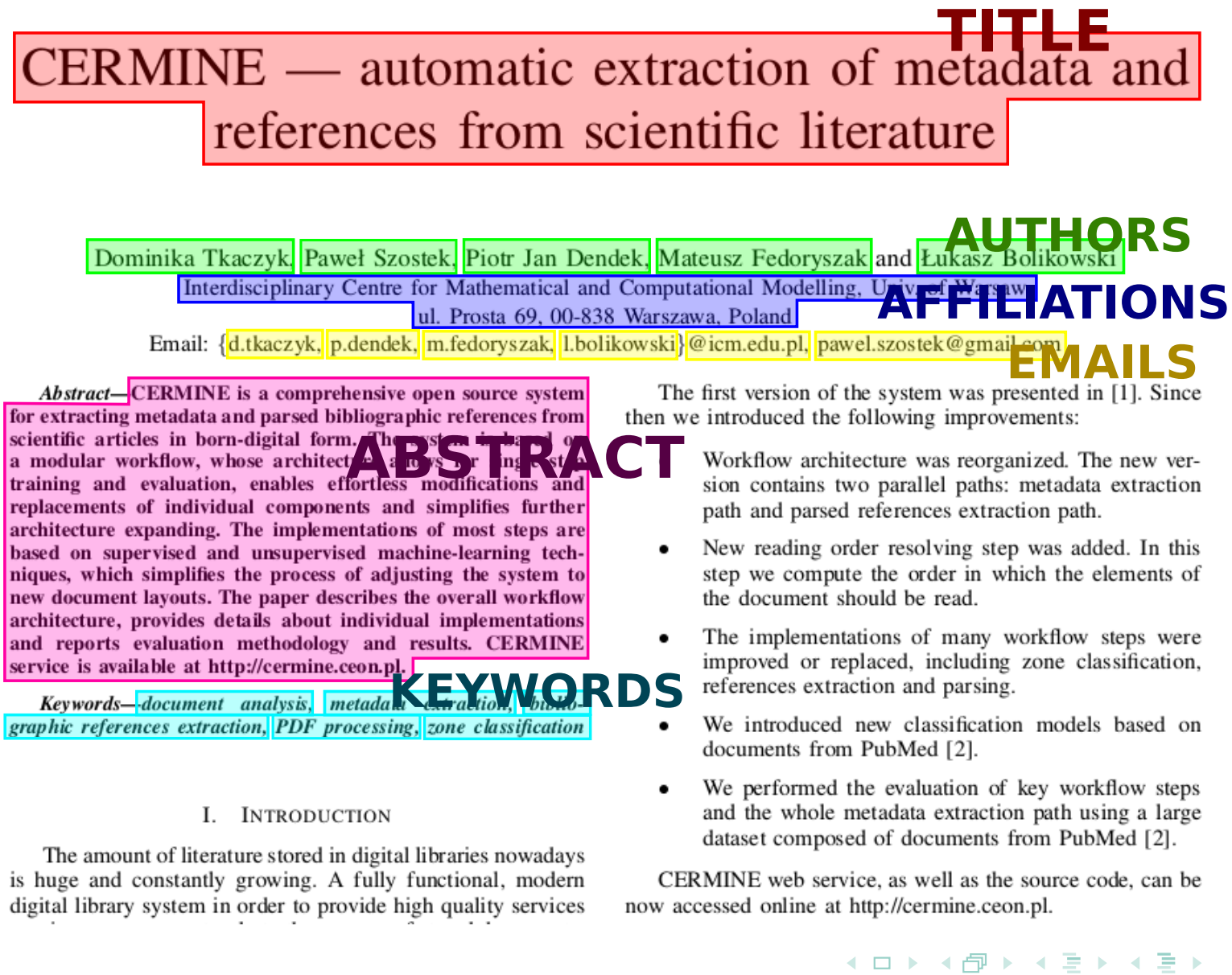
CERMINE — automatic extraction of metadata and references from scientific literature

Dominika Tkaczyk, Pawel Szostek,
Piotr Jan Dendek, Mateusz Fedoryszak and Lukasz Bolikowski

Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw

11th IAPR International Workshop on Document Analysis Systems
7-10 April 2014

The goal



TITLE

CERMINE — automatic extraction of metadata and references from scientific literature

AUTHORS

Dominika Tkaczyk, Paweł Szostek, Piotr Jan Dendek, Mateusz Fedoryszak and Łukasz Bolikowski

AFFILIATIONS

Interdisciplinary Centre for Mathematical and Computational Modelling, Univ. of Warsaw
ul. Prosta 69, 00-838 Warszawa, Poland

EMAILS

Email: {d.tkaczyk, p.dendek, m.fedoryszak, l.bolikowski}@icm.edu.pl, pawel.szostek@gmail.com

ABSTRACT

***Abstract*—CERMINE is a comprehensive open source system for extracting metadata and parsed bibliographic references from scientific articles in born-digital form. The system is based on a modular workflow, whose architecture allows training and evaluation, enables effortless modifications and replacements of individual components and simplifies further architecture expanding. The implementations of most steps are based on supervised and unsupervised machine-learning techniques, which simplifies the process of adjusting the system to new document layouts. The paper describes the overall workflow architecture, provides details about individual implementations and reports evaluation methodology and results. CERMINE service is available at http://cermine.ceon.pl.**

KEYWORDS

***Keywords*—document analysis, metadata extraction, bibliographic references extraction, PDF processing, zone classification**

I. Introduction

The amount of literature stored in digital libraries nowadays is huge and constantly growing. A fully functional, modern digital library system in order to provide high quality services

The first version of the system was presented in [1]. Since then we introduced the following improvements:

- Workflow architecture was reorganized. The new version contains two parallel paths: metadata extraction path and parsed references extraction path.
- New reading order resolving step was added. In this step we compute the order in which the elements of the document should be read.
- The implementations of many workflow steps were improved or replaced, including zone classification, references extraction and parsing.
- We introduced new classification models based on documents from PubMed [2].
- We performed the evaluation of key workflow steps and the whole metadata extraction path using a large dataset composed of documents from PubMed [2].

CERMINE web service, as well as the source code, can be now accessed online at http://cermine.ceon.pl.

The goal

- performing the evaluation of the whole references extraction path using the PubMed-based dataset,
- the evaluation of other similar systems using the same dataset and comparing the extraction results.

ACKNOWLEDGMENTS

The work is supported by the National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10 by the Strategic scientific research and experimental development program: "Interdisciplinary System for Interactive Scientific and Scientific-Technical Information".

REFERENCES

[1] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek, "A modular metadata extraction system for born-digital articles," in *10th IAPR International Workshop on Document Analysis Systems*, 2012, pp. 11–16.

[2] "PubMed," http://www.ncbi.nlm.nih.gov/pubmed.

[3] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," in *3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2003, pp. 37–48.

[4] S. Marinai, "Metadata Extraction from PDF Papers for Digital Library Ingest," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 251–255.

Conference on Digital Libraries, 2009, pp. 73–76.

[9] R. Kern, K. Jack, and M. Hristakeva, "TeamBeam - Meta-Data Extraction from Scientific Literature," *D-Lib Magazine*, vol. 18, 2012.

[10] A. Constantin, S. Pettifer, and A. Voronkov, "Pdfx: fully-automated pdf-to-xml conversion of scientific literature," in *2013 ACM Symposium on Document Engineering*, 2013, pp. 177–180.

[11] "NLM," http://dtd.nlm.nih.gov/archiving/.

[12] C. H. Lee and T. Kanungo, "The architecture of TrueViz: a groundTRUth/metadata editing and VIsualiZing ToolKit," *Pattern Recognition*, vol. 15, 2002.

[13] "PdfMiner," http://www.unixuser.org/ euske/python/pdfminer/.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[15] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002.

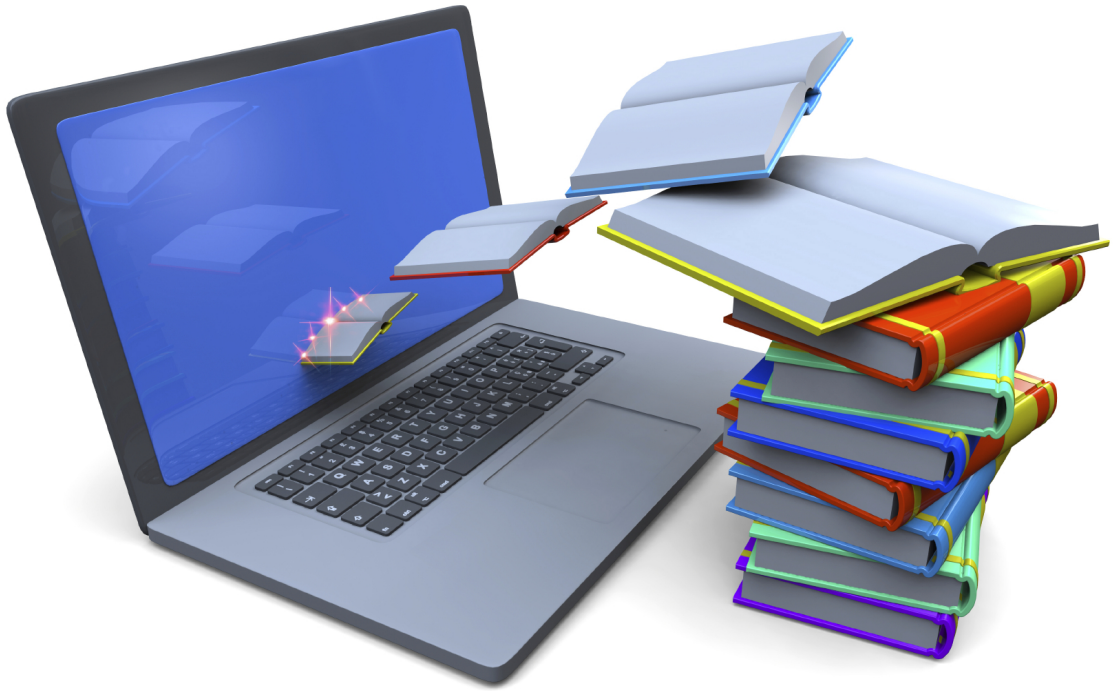
[16] D. Tkaczyk, A. Czeczko, K. Rusek, L. Bolikowski, and R. Bogacewicz, "Grotoap: ground truth for open access publications," in *12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2012, pp. 381–382.

[17] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *3rd ACM Conference on Digital Libraries*. ACM, 1998, pp. 89–98.

[18] A. McCallum, K. Nigam, and J. Rennie, "Automating the construction of internet portals with machine learning," *Information Retrieval*, pp. 127–163, 2000.

AUTHOR URL SOURCE TITLE PAGES VOLUME YEAR

The motivation



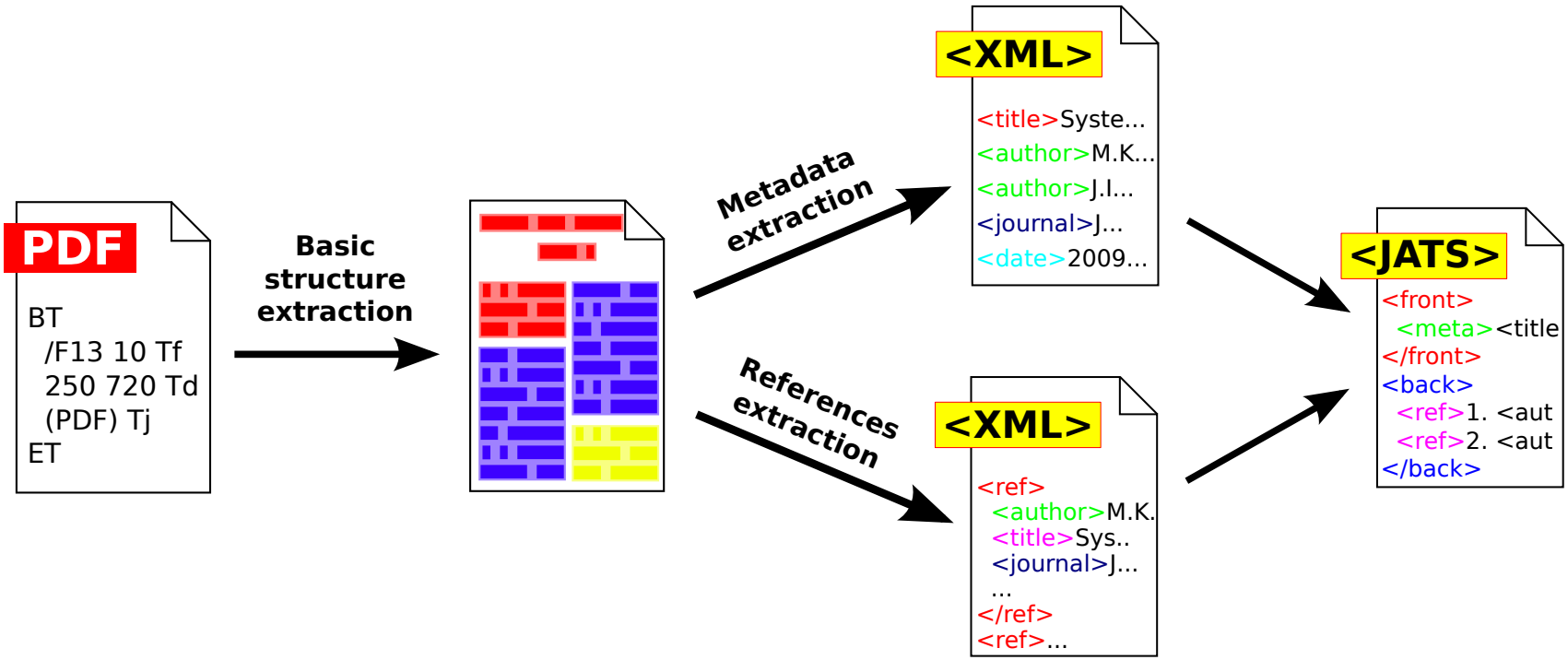
- There are **documents without metadata**.
- **Metadata information** may be **incomplete** or **incorrect**.

Requirements

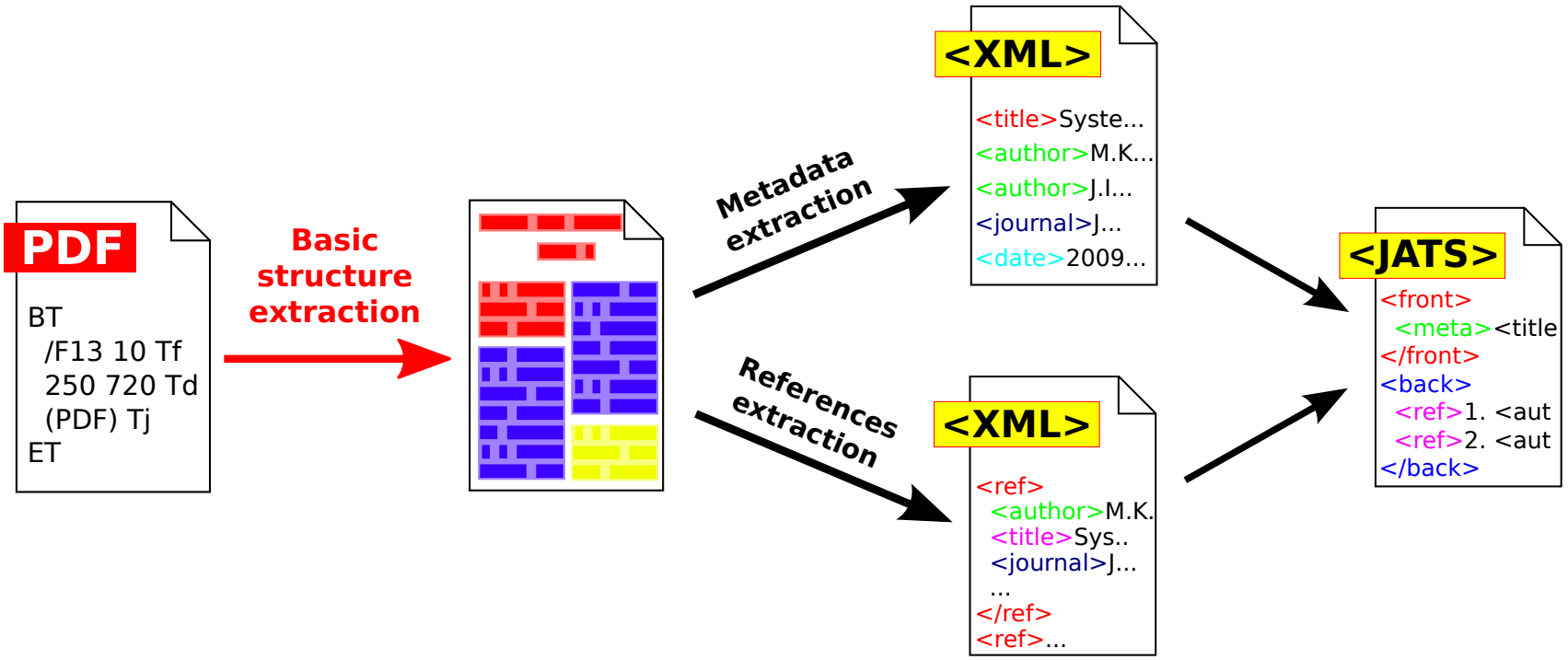
The **metadata extraction system** should be:

- **comprehensive**,
- **automatic**,
- **modular**,
- **open** and **widely available**,
- **easily applicable**,
- **flexible** and able to **adapt to new layouts**,
- **well tested**.

The process

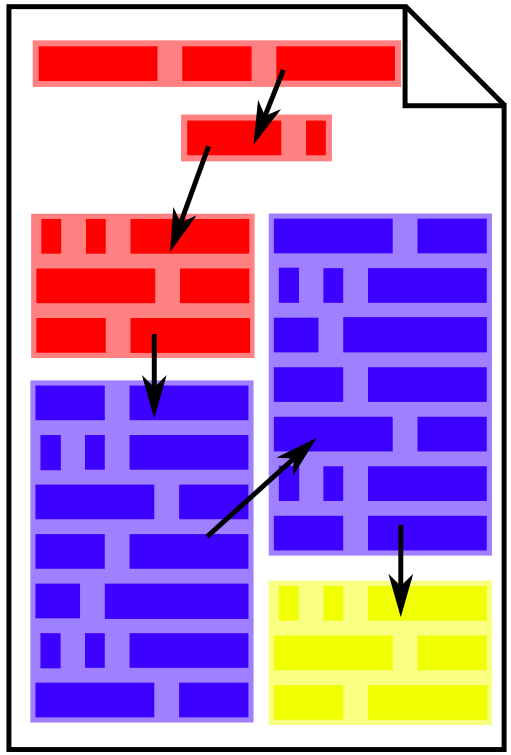

PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET
Basic structure extraction
Metadata extraction
References extraction
<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
<XML>
<ref>
<author>M.K.
<title>Sys..
<journal>J...
...
</ref>
<ref>...
<JATS>
<front>
<meta><title
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

The process


PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET
Basic structure extraction
Metadata extraction
References extraction
<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
<XML>
<ref>
<author>M.K.
<title>Sys..
<journal>J...
...
</ref>
<ref>...
<JATS>
<front>
<meta><title
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

Basic structure extraction

- Character extraction — **iText** library
- Page segmentation — **Docstrum**
- Reading order resolving — bottom-up **heuristic-based**
- Initial zone classification — **SVM** (*metadata*, *references*, *body* and *other*)



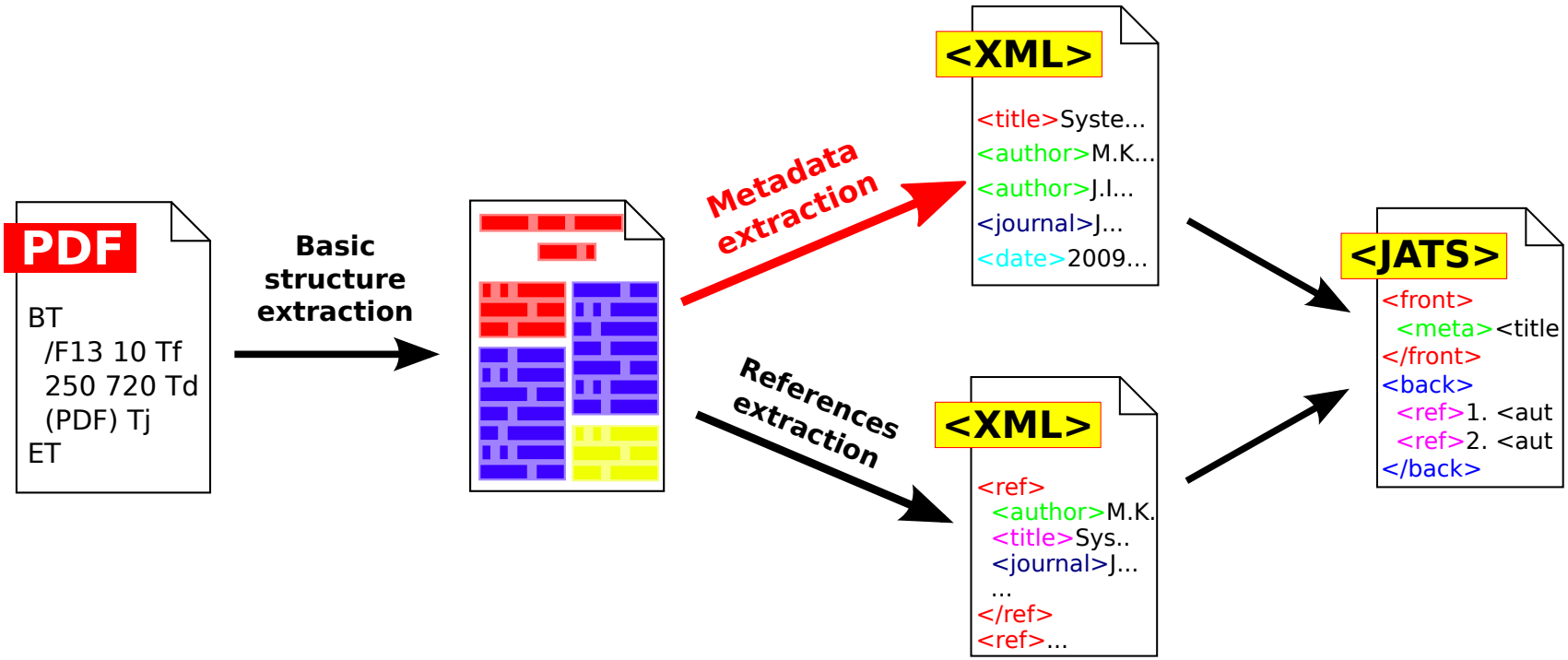
The output

TrueViz XML format:

- **hierarchical structure** containing: pages, zones, lines, words, characters
- all elements have **bounding boxes**
- **reading order** is given
- zones have **labels**

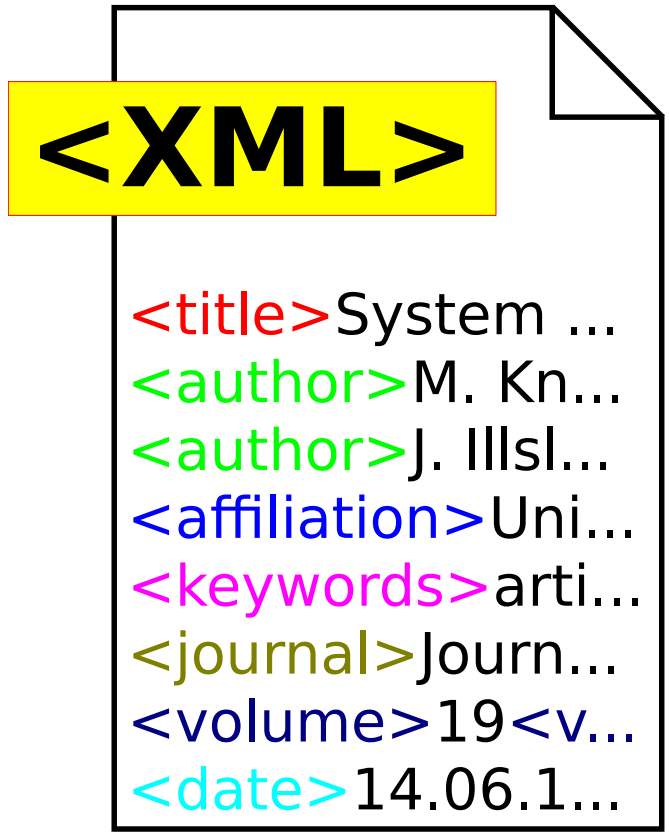
```
<Page>
<PageID Value="0"/>
<Zone>
<ZoneID Value="0"/>
<ZoneCorners>
<Vertex x="55.320"y="34.295"/>
<Vertex x="235.704"y="58.295"/>
</ZoneCorners>
<ZoneNext Value="1"/>
<Category Value="TITLE"/>
<Line>
<Word>
<Character>
```

The process


PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET
Basic structure extraction
Metadata extraction
References extraction
<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
<XML>
<ref>
<author>M.K.
<title>Sys..
<journal>J...
...
</ref>
<ref>...
<JATS>
<front>
<meta><title
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

Metadata extraction

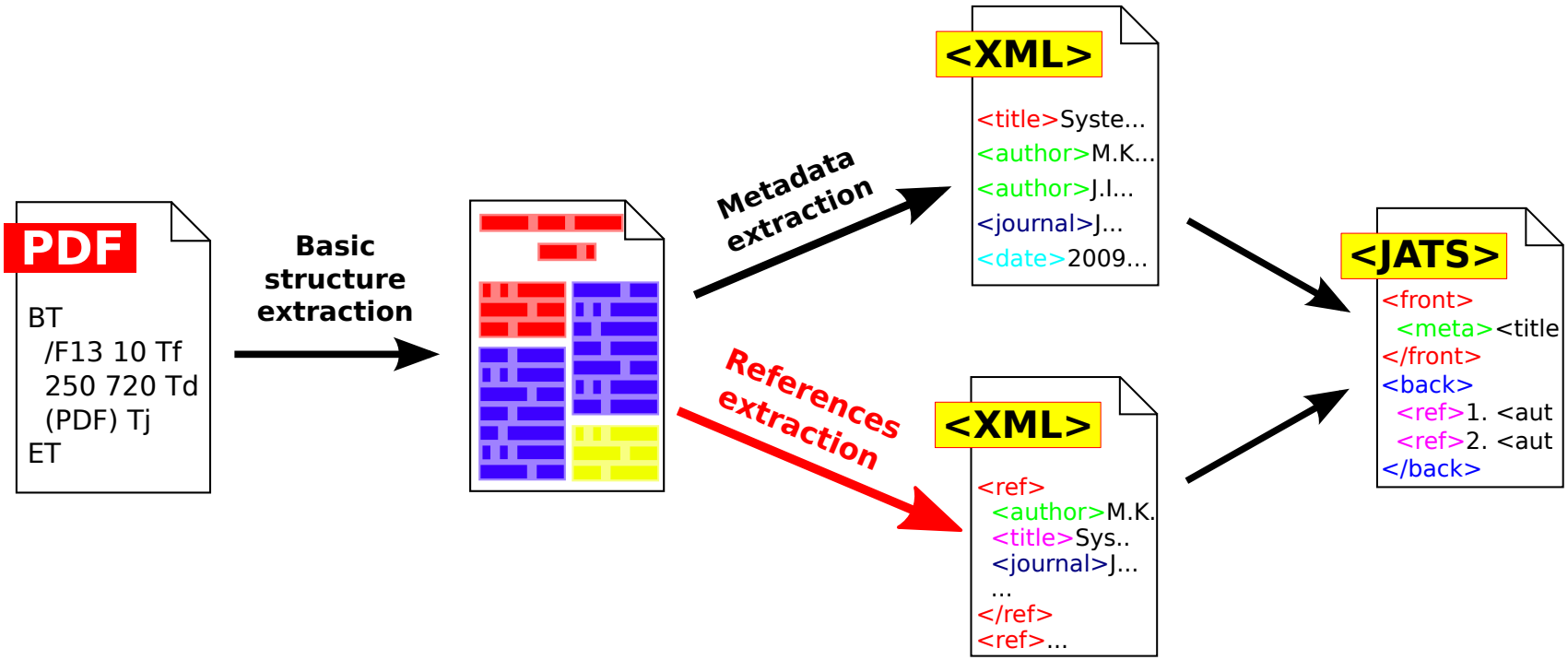
- Metadata zone classification — **SVM** (*abstract*, *bib_info*, *type*, *title*, *affiliation*, *author*, *keywords*, *correspondence*, *dates* and *editor*)
- Metadata extraction — simple **rule-based**



Zone classification

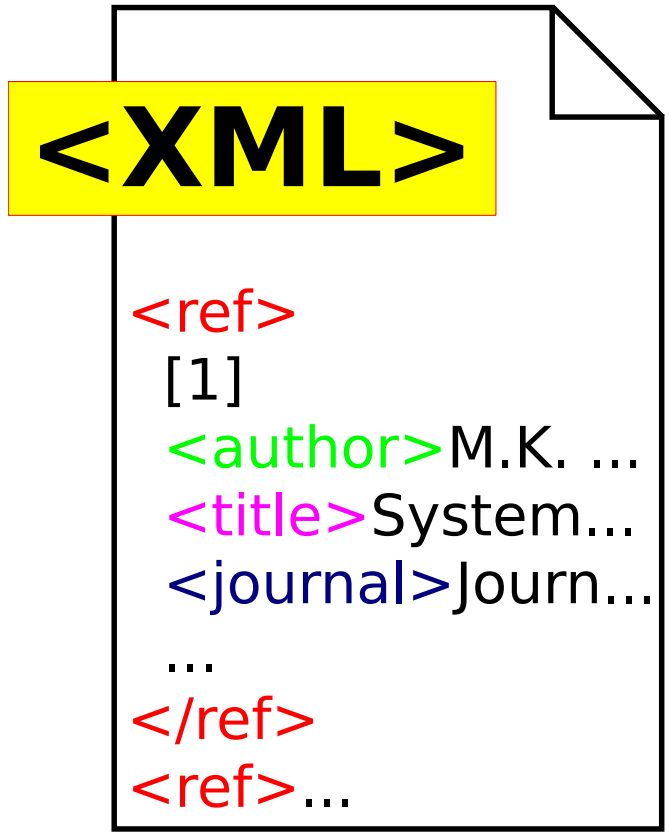
- classifiers are based on **LibSVM** library
- a zone is represented by **78 features**: **geometrical**, **lexical**, **sequential**, **formatting**, **heuristics**
- the best **SVM parameters** were found by:
  - a **grid-search** over 3-dimensional space of kernel function types and C (penalty parameter) and γ coefficients
  - at every grid point a **10-fold cross-validation** was performed
  - we chose the parameters that gave the best **mean accuracy**
- initial classifier was trained on **964 documents** with **155,144 zones** in total
- metadata classifier was trained on **1,934 documents** and **45,035 metadata zones** in total

The process

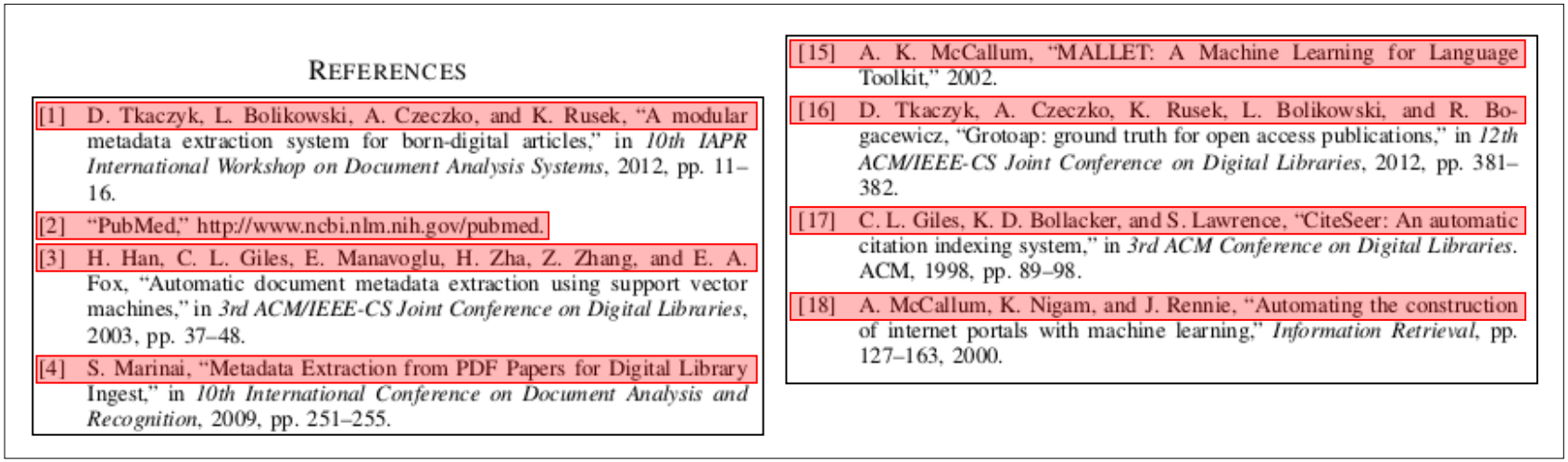

PDF
BT
/F13 10 Tf
250 720 Td
(PDF) Tj
ET
Basic structure extraction
Metadata extraction
References extraction
<XML>
<title>Syste...
<author>M.K...
<author>J.I...
<journal>J...
<date>2009...
<XML>
<ref>
<author>M.K.
<title>Sys..
<journal>J...
...
</ref>
<ref>...
<JATS>
<front>
<meta><title
</front>
<back>
<ref>1. <aut
<ref>2. <aut
</back>

Parsed reference extraction

- Reference strings extraction — **K-means clustering**
- Reference parsing — **CRF**



Reference strings extraction



REFERENCES

[1] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek, "A modular metadata extraction system for born-digital articles," in *10th IAPR International Workshop on Document Analysis Systems*, 2012, pp. 11–16.

[2] "PubMed," http://www.ncbi.nlm.nih.gov/pubmed.

[3] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox, "Automatic document metadata extraction using support vector machines," in *3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2003, pp. 37–48.

[4] S. Marinai, "Metadata Extraction from PDF Papers for Digital Library Ingest," in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 251–255.

[15] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," 2002.

[16] D. Tkaczyk, A. Czeczko, K. Rusek, L. Bolikowski, and R. Bogacewicz, "Grotoap: ground truth for open access publications," in *12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2012, pp. 381–382.

[17] C. L. Giles, K. D. Bollacker, and S. Lawrence, "CiteSeer: An automatic citation indexing system," in *3rd ACM Conference on Digital Libraries*. ACM, 1998, pp. 89–98.

[18] A. McCallum, K. Nigam, and J. Rennie, "Automating the construction of internet portals with machine learning," *Information Retrieval*, pp. 127–163, 2000.

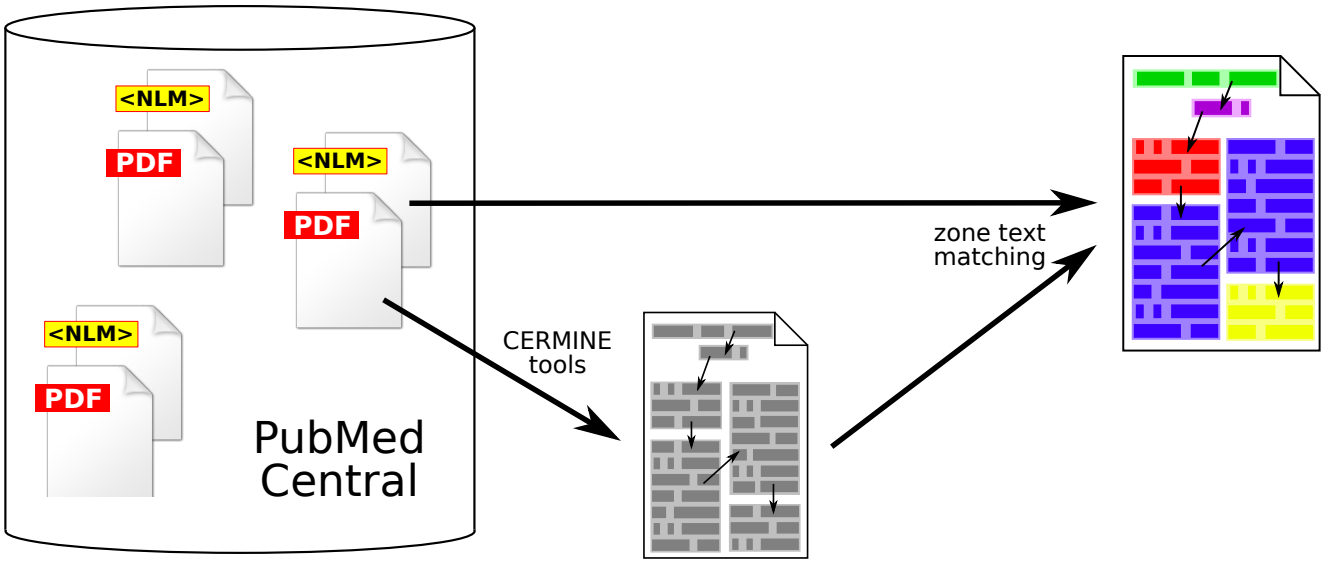
- **clustering text lines into two sets**: first lines and the rest
- unsupervised **K-means algorithm** with **Euclidean distance**
- **5 features** (based on length, indentation, space between lines and the text)

Reference parsing

[8] Y. Wang, I.T. Phillips and R.M. Haralick, Document zone content classification and its performance evaluation, Pattern Recognition 39 (1) (2006), pp. 57–73.

- **Conditional Random Fields** token classifier based on **GRMM** and **MALLET** packages
- **42 constant features** + the most popular **words** + features of **two preceding** and **two following** tokens
- the classifier was trained on **1000 citations** from **Cora-ref** + **PubMed**

GROTOAP2 dataset



- **GROund Truth for Open Access Publications**
- built automatically from **PubMed Central Open Access Subset**
- $\sim$ **60k ground truth files** in TrueViz format with corresponding PDF files

Results

	avg. precision	avg. recall
initial zone classifier	91.74%	87.31%
metadata zone classifier	92.49%	93.83%
reference parsing	90.18%	89.51%

	precision	recall
journal title	68.68%	49.23%
volume	97.57%	78.57%
issue	52.50%	56.64%
pages	51.37%	34.71%
year	98.79%	89.18%
DOI	93.60%	57.46%
ISSN	44.29%	3.01%

	avg. adjustment
article title	95.03%
abstract	91.43%

	avg. precision	avg. recall
authors	87.19%	82.07%
affiliations	70.13%	59.44%
keywords	61.11%	68.37%

Future work

- a new extraction path for **extracting structured full text**
- the **evaluation** of the entire **references extraction path**
- **comparing the results** to other similar systems

Links

- **CERMINE web service**: `http://cermine.ceon.pl`
- **CERMINE source code**: `https://github.com/CeON/CERMINE`
- **GROTOAP2**: `http://cermine.ceon.pl/grotoap2/`

Thank you!
Questions?

Dominika Tkaczyk
d.tkaczyk@icm.edu.pl