Back to meaning - information structuring in the PEER project

Foudil Bretel, Patrice Lopez, Maud Medves, Alain Monteil, Laurent Romary

Back to meaning – information structuring in the PEER project

Foudil Bretel, Patrice Lopez, Maud Medves, Alain Monteil, Laurent Romary

INRIA & HUB IDSL

One of the major challenges in textual data management is certainly to deal with the huge legacy of unstructured documents, which, whether created by Text processor (MSWord) or PDFs, have populated our hard disks and online repositories over the years. Indeed, most of these documents hardly give out any clue for further exploitation: whether metadata or, all the more, full-text content, they are basically doomed to be retrieved as just bags of words (often of characters for PDFs) until a human eye manages to give actual meaning to it.

This situation has been made even more complex in the publishing world, where, for lack of enough dedicated technical resources and actual coordination, publishers are producing metadata information associated with (mostly PDF) scholarly documents in a wide variety of formats, which can be seen as a major informational chaos. This is all the more annoying since any further use of a scholarly paper can only be based on the precise identification of, among other things, the publication and author details.

In this context we would like to present in this paper a specific attempt to sort out part of this chaos in the context of the EU PEER project. Initiated as an experiment to observe the consequence of large scale author manuscript deposit in publication repositories, the project led to the design and implementation of an information HUB (the *PEER Depot*) where all publishers' data (the author's manuscript and, when available, publishers' metadata) would be normalized so that they could be further uploaded to a series of trusted publication repositories to be put eventually in open access.

To do so, we actually used the TEI as the *lingua franca* for all metadata exchange within this workflow, seeing already this information as the seed of full featured TEI document (<biblStruct> in <sourceDesc>, abstract as a in <front>), in case the technology allows us to get more information from the documents in the future. In our presentation we will report on two major activities that we carried out within the PEER Depot:

- The mapping of a large variety of publishers' data (NLM 2.x, NLM 3.0, ScholarOne, proprietary (Springer, Elsevier, Nature, Sage, BMJ, …)) onto a single and well constrained TEI structure. In particular, this activity has led to the identification of precise encoding "best practices" concerning bibliographical or affiliation information, implemented as a series of large coverage stylesheets;

- A more prospective action on extracting information automatically from a pdf file in the case when wewanted to obtain additional metadata. To this end we used the GROBID[1] environment and trained it to match various title page styles in scholarly papers. As exemplified in the annex below we obtained particularly good results

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

allowing us to automate the integration of pdfs within the PEER depot for all documents provided by some publishers.

At this stage, we can already conclude that the TEI infrastructure, with both its expressivity and its customization capabilities has been a perfect choice in the project. Even more, it allowed us to point to an essential issue: if we are to think of long term accessibility to scholarly documents, this should be done in a wider context of textual encoding rather than within a niche (like the one provided by the NLM format(s)) that may just be more difficult to maintain.

On the technological side, the first experiments with automatic extraction of structured information from texts with GROBID have made us very optimistic about the prospect of generalizing this to the full textual content of scholarly papers available in PDFs. Automatic extraction can be highly valuable for tasks such as bibliometrics or information retrieval, as it provides additional information (affiliation, address, keywords, abstract) to publishers' metadata.

This in turns gives hope for defining long-term archiving and exploitation strategies for such documents.

Annex I: Front page of a scholarly article

Time resolved optical emission spectroscopy of an HPPMS coating process

S Theiß[1], N Bibinov[2], N Bagcivan[1], M Ewering[1], P Awakowicz[2] and K Bobzin[1]

[1] Surface Engineering Institute, RWTH Aachen University, Augustinerbach 4-22, D-52062 Aachen, Germany

[2] Institute for Electrical Engineering and Plasma Technology, Ruhr-Universität Bochum, D-44780 Bochum, Germany

E-mail: theiss@iot.rwth-aachen.de

Abstract. This paper deals with the time resolved optical emission spectroscopy (OES) of a high power pulse magnetron sputtering (HPPMS) physical vapor deposition (PVD) coating process. With an industrial coating unit CC800/9 HPPMS (CemeCon AG, Würselen) a (Cr,Al,Si)N coating was deposited. During the coating process, an absolute calibrated Echelle-spectrometer (ESA-3000) measured the intensities of the spectral lines of Chromium (Cr), Aluminum (Al) and molecular bands of nitrogen (N_2). Time resolved measurements enable us to calculate different parameters like the average velocity of sputtered Al- and Cr-atoms or the internal plasma parameters electron density n_e and electron temperature kT_e with a time resolution of 20 μs. With these parameters, we determine the ionization rates of Al-, Cr-, Ar- and Kr-atoms and the deposition densities of Al- and Cr-atoms. Thus simulated deposition densities of $1.75 \cdot 10^{20}$ $m^{-2}s^{-1}$ for Chromium and $1.7 \cdot 10^{22}$ $m^{-2}s^{-1}$ for Aluminum are reached.

Keywords: HPPMS, HIPIMS, OES, optical emission spectroscopy, absolute calibrated
Submitted to: *J. Phys. D: Appl. Phys.*

Annex II: Automatically extracted information from the article


```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="../jsp/xmlverbatimwrapper.xsl"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xlink="http://www.w3.org/1999/xlink"
 xmlns:mml="http://www.w3.org/1998/Math/MathML">
    <teiHeader xml:lang="en">
        <fileDesc>
            <titleStmt>
                <title level="a">Time resolved optical emission spectroscopy of an HPPMS coating
process</title>
            </titleStmt>
            <publicationStmt>
                <date>2010</date>
            </publicationStmt>
            <sourceDesc>
                <biblStruct>
                    <analytic>
                        <author role="corresp" type="corresp">
                            <persName>
                                <forename type="first">S</forename>
                                <surname>Theiß</surname>
                            </persName>
                            <email>theiss@iot.rwth-aachen.de</email>
                            <affiliation>
                                <orgName type="department">Surface Engineering Institute</orgName>
                                <orgName type="institution">RWTH Aachen University</orgName>
                                <address>
                                    <addrLine>Augustinerbach 4-22</addrLine>
                                    <postCode>D-52062</postCode>
                                    <settlement>Aachen</settlement>
                                    <country key="DE">Germany</country>
                                </address>
                            </affiliation>
                        </author>
                        <author>
                            <persName>
                                <forename type="first">N</forename>
                                <surname>Bibinov</surname>
                            </persName>
                            <affiliation>
                                <orgName type="department">Institute for Electrical Engineering and
Plasma Technology</orgName>
                                <orgName type="institution">Ruhr-Universität Bochum</orgName>
                                <address>
                                    <postCode>D-44780</postCode>
                                    <settlement>Bochum</settlement>
                                    <country key="DE">Germany</country>
                                </address>
                            </affiliation>
                        </author>
                        <author>
                            <persName>
```

```
                <forename type="first">N</forename>
                <surname>Bagcivan</surname>
            </persName>
            <affiliation>
                <orgName type="department">Surface Engineering Institute</orgName>
                <orgName type="institution">RWTH Aachen University</orgName>
                <address>
                    <addrLine>Augustinerbach 4-22</addrLine>
                    <postCode>D-52062</postCode>
                    <settlement>Aachen</settlement>
                    <country key="DE">Germany</country>
                </address>
            </affiliation>
        </author>
        <author>
            <persName>
                <forename type="first">M</forename>
                <surname>Ewering</surname>
            </persName>
            <affiliation>
                <orgName type="department">Surface Engineering Institute</orgName>
                <orgName type="institution">RWTH Aachen University</orgName>
                <address>
                    <addrLine>Augustinerbach 4-22</addrLine>
                    <postCode>D-52062</postCode>
                    <settlement>Aachen</settlement>
                    <country key="DE">Germany</country>
                </address>
            </affiliation>
        </author>
        <author>
            <persName>
                <forename type="first">P</forename>
                <surname>Awakowicz</surname>
            </persName>
            <affiliation>
                <orgName type="department">Institute for Electrical Engineering and
Plasma Technology</orgName>
                <orgName type="institution">Ruhr-Universität Bochum</orgName>
                <address>
                    <postCode>D-44780</postCode>
                    <settlement>Bochum</settlement>
                    <country key="DE">Germany</country>
                </address>
            </affiliation>
        </author>
        <author>
            <persName>
                <forename type="first">K</forename>
                <surname>Bobzin</surname>
            </persName>
            <affiliation>
                <orgName type="department">Surface Engineering Institute</orgName>
                <orgName type="institution">RWTH Aachen University</orgName>
                <address>
```

```
                                <addrLine>Augustinerbach 4-22</addrLine>
                                <postCode>D-52062</postCode>
                                <settlement>Aachen</settlement>
                                <country key="DE">Germany</country>
                            </address>
                        </affiliation>
                    </author>
                    <title level="a">Time resolved optical emission spectroscopy of an HPPMS coating process</title>
                </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">7</biblScope>
                        <biblScope type="fpage">75205</biblScope>
                        <date>2010</date>
                    </imprint>
                </monogr>
                <idno type="doi">10.1088/0022-3727/43/7/075205</idno>
                <note type="submission">Submitted to : J . Phys . D : Appl . Phys .</note>
                <note>Time. resolved optical emission spectroscopy of an HPPMS coating process 2</note>
                <keywords>HPPMS, HIPIMS, OES, optical emission spectroscopy, absolute calibrated</keywords>
            </biblStruct>
        </sourceDesc>
    </fileDesc>
</teiHeader>
<text xml:lang="en">
    <front>
        <div type="abstract">
            <head>Abstract</head>
<p>This paper deals with the time resolved optical emission spectroscopy (OES) of a high power pulse magnetron sputtering (HPPMS) physical vapor deposition (PVD) coating process. With an industrial coating unit CC800/9 HPPMS (CemeCon AG, Würselen) a (Cr, Al, Si)N coating was deposited. During the coating process, an absolute calibrated Echelle-spectrometer (ESA-3000) measured the intensities of the spectral lines of Chromium (Cr), Aluminum (Al) and molecular bands of nitrogen (N 2). Time resolved measurements enable us to calculate different parameters like the average velocity of sputtered Al- and Cr-atoms or the internal plasma parameters electron density n e and electron temperature kT e with a time resolution of 20 µs. With these parameters, we determine the ionization rates of Al-, Cr-, Ar- and Kr-atoms and the deposition densities of Al- and Cr-atoms. Thus simulated deposition densities of 1.75 · 10 20 m −2 s −1 for Chromium and 1.7 · 10 22 m −2 s −1 for Aluminum are reached.</p>
        </div>
    </front>
</text>
</TEI>
```