TEI-based representation of the PEER metadata profile

Laurent Romary, Maud Medves

TEI-based representation of the PEER metadata profile[1]

Laurent Romary & Maud Medves

INRIA & HUB IDSL

Introduction

In the context of rapid growth of institutional repositories, the PEER project (www.peerproject.eu) has been set up to monitor the effects of systematic archiving of author's final peer-reviewed manuscript over time.

To conduct this research, interoperability between editor's data and online repositories is an essential asset. It is therefore important to specify the encoding scheme of required metadata elements, particularly the ones related to author's identification (as it is particularly important for citations, e.g. databases such as the ones accessed through Web of Science).

A high granularity of metadata is preferable, as it allows for deeper structured information and enables precise information to be identified and shared. It offers indeed greater flexibility in treating each data field in isolation if required.

The PEER project has adopted the Text Encoding Initiative encoding scheme (www.tei-c.org).

This document describes the output interchange that was tuned for the PEER Depot for the integration and distribution of the metadata information provided by publishers within the PEER project. It is based on the TEI's Guidelines for Text Encoding and Interchange (P5), with some additional constraints intended to make the corresponding information structures univocally interpretable.

Overview

The proposed structure combines a global structure (<TEI>[2]), which can potentially integrate any information that can be found in a full-text representation of a paper article, and a sub-structure (<biblStruct>[3]) that specifically contains the bibliographical information of the article. This allows us to process in a uniform way the two following scenarios:

[1] Metadata profile adopted within the PEER project (ECP-2007-DILI-537003)

[2] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-TEI.html

[3] http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-biblStruct.html

- The PEER Depot receives full-text articles in XML (or retrieves them from repositories such as PMC) and converts them to the TEI format, thus exploiting all its expressive capacities.
- The PEER Depot receives specific metadata information, with possibly some additional content (e.g. abstract). A highly simplified <TEI> structure is created, which is mainly a container for disseminating the bibliographical content.

The remaining part of this document will primarily address the second scenario, which is the one needed for the research to be carried out within the PEER project.

General structure of a TEI document

The TEI information model is intended to represent both the textual content of a document and the metadata attached to it. This is reflected in the two main parts of a <TEI> root element, namely <TEIHeader> and <text>.

The TEI header is in turn organised in a series of sub-components:

- <fileDesc> gathering the main characteristics of the document (title, author, bibliographic description of the source)
- <profileDesc> providing some information about the content (e.g. languages used in the text, keywords)
- <revisionDesc> providing the history of the document

The <text> element is further decomposed in <front>, <body> and <back>. Where available, abstracts are represented in <front> and full-text content in subsequent elements.

Skeleton of a full TEI document (as relevant for PEER) for the representation of scholarly articles

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title level="a" type="main">...</title>
            </titleStmt>
            <publicationStmt>
                <availability>
                    <p>Copyright © The Animal Consortium 2009</p>
                </availability>
                <date>2009</date>
                <authority>The Animal Consortium</authority>
            </publicationStmt>
            <sourceDesc>
                <biblStruct>...</biblStruct>
            </sourceDesc>
        </fileDesc>
        <profileDesc>
            <textClass>
                <keywords>
                    <list>
```

```
                        <head>Keywords</head>
                        <item>
                            <term>foetal development</term>
                        </item>
                        <item>
                            ...
                        </item>
                    </list>
                </keywords>
            </textClass>
        </profileDesc>
        <revisionDesc>
            <change when="2008-08-27">Received</change>
            <change when="2008-12-01">Accepted</change>
        </revisionDesc>
    </teiHeader>
    <text>
        <front>
            <div type="abstract">
                <head>Abstract</head>
                <div>
                <p>...</p>
                </div>
            </div>
        </front>
        <body/>
        <back/>
    </text>
</TEI>
```

Representation of bibliographical information

The representation is based on the TEI <biblStruct> element, which is organised as follows:

```
<biblStruct type="article">
    <analytic>
        …
    </analytic>
    <monogr>
        …
        <imprint>
            …
        </imprint>
    </monogr>
    …
</biblStruct>
```

A <biblStruct> is mainly divided into two sub-structures:

- <analytic> to indicate the bibliographical characteristics of an article (title and authors)
- <monogr> to account for the publication details of the journal (journal name, publisher information, ISSN, etc.). It contains in turn an <imprint> element which gathers publication and/or distribution aspects of the article in the corresponding journal (pagination, volume, issue, etc.)

When applicable, additional notes or identifiers can follow, for instance, the DOI, pubmed-central-id or repository-specific-id will appear here:

```
<biblStruct type="article">
    <analytic>…</analytic>
    <monogr>…</monogr>
    <idno type="pmid">12345678</idno>
</biblStruct>
```

The <analytic> element

The title of a journal article is represented by means of the <title> element (with appropriate @level attribute) as follows:

```
<title level="a">Multilocus Analysis of Age Related Macular
Degeneration</title>
```

When necessary, a further @type attribute may be used to differentiate between main and subtitles (@type="main" vs. @type="sub").

Each author in the <analytic> element is independently described by means of an <author> element. This element contains the author's name, affiliation and addresses – when available – as presented in the outline below:

```
<author>
     <idno type="...">...</idno>
     <persName>
        <forename>Michael</forename>
        <surname>Dean</surname>
     </persName>
     <affiliation>…</affiliation>
     <email>dean@ncifcrf.gov</email>
</author>
```

For more details about encoding the affiliation element, see the chapter "Dealing with affiliation" below.

The <monogr> element

The <monogr> element gathers journal identification information (journal title and ISSN) together with the publishing information contained in its <imprint> sub-element, for instance:

```
<monogr>
   <title level="j" type="main">European Journal of Human Genetics</title>
   <title level="j" type="nlm-ta">Eur J Hum Genet</title>
   <idno type="ISSN">1018-4813</idno>
   <imprint>…</imprint>
</monogr>
```

The <imprint> element

“By imprint is meant all the information relating to the publication of a work: the person or organization by whose authority and in whose name a bibliographic entity such as a book is made public or distributed (whether a commercial publisher or some other organization), the place of publication, and a date. It may also include a full address for the publisher or organization. Full bibliographic references usually specify either the number of pages in a print publication (or equivalent information for non-print materials), or the specific location of the material being cited within its containing publication.”[4]

The <imprint> element is organised as follows:

```
<imprint>
  <pubPlace>Oxford</pubPlace>
  <publisher>Clarendon Press</publisher>
  <date type=”Published” when="1969-02-07"/>
  <biblScope type="vol">3</biblScope>
  <biblScope type="issue">2</biblScope>
</imprint>
```

The possible values for the attribute type on <biblScope> are the following:

- vol: volume
- issue: issue
- fpage: first page
- lpage: last page
- pp: number of pages when the information about full pagination is not available[5]

The possible values for the attribute type on <date> are the following:

- Published: date of publication (online or imprint - when no further precision is available)[6]
- ePublished: date of online publication (when available)
- pPublished: date of imprint publication (when available)

<biblStruct> skeleton

The following example provides an overview of the full internal structure of the <biblStruct> element as provided by the PEER Depot within a <TEI> document. Most mandatory PEER metadata fields are illustrated here.

```
<biblStruct type="article">
    <analytic>
        <title level="a" type="main">…</title>
        <author type="corresp">
            <persName>
```

[4] http://www.stoa.org/projects/epidoc/stable/guidelines/

[5] We restrict here the semantic of the recommended value (cf. http://www.tei-c.org/release/doc/tei-p5-doc/html/ref-biblScope.html).

[6] Default value - Published is used when no further precision about online and imprint publication are associated to a given date.

```
                <forename>…</forename>
                <surname>…</surname>
            </persName>
            <affiliation>
                <orgName type="">…</orgName>
                <address>
                  …
                  <country key="FR">France</country>
               </address>
            </affiliation>
            <email>…</email>
        </author>
    </analytic>
    <monogr>
        <title level="j" type="main">…</title>
        <idno type="ISSN">…</idno>
        <imprint>
            <publisher>…</publisher>
            <pubPlace>…</pubPlace>
            <date when="2009-02-03"/>
            <biblScope type="fpage">…</biblScope>
        </imprint>
    </monogr>
    <idno type="DOI">…</idno>
</biblStruct>
```

Mapping table

The following table makes explicit the PEER mandatory metadata fields, as found in the TEI based exchange format transferred to PEER repositories.

Field Name	**Path in TEI document**	**Notes**
Article title	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/analytic/title[@type='main']	When applicable additional titles are provided (with specific values of @type).
Correspond-ing author	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/analytic/author[@type='corresp']	Additional authors are provided as siblings of this element in further <author> elements.
Author name	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/analytic/author/persName	The following elements are used for describing author's name: <forename>, <surname>, <roleName>, <nameLink>, <genName>
Author email	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/analytic/author/email	
Abstract	/TEI/text/front/div[@type='abstract']	Further elements may be found in the abstract, most notably: <head> for abstract title <p> for paragraphs <hi> for additional rendering (e.g. <hi rend="italic">)
Publication date	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/monogr/imprint/date/@when	Expressed in conformance to ISO 8601:2004 (i.e. yyyy-MM-dd)
DOI of published article	/TEI/teiHeader/fileDesc/sourceDesc/ biblStruct/idno[@type='DOI']	When applicable, further identifiers maybe provided with additional <idno> elements.

Country of contributing authors	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/analytic/author/affiliation/address/country/@key	Expressed in conformance to ISO 3166-1-A2 (e.g. FR).
Journal title	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/title[@type=’main’]	Additional titles (e.g. abbreviated) may appear with publisher specific @type values.
Affiliation	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/analytic/author/affiliation	Main components here are expressed in <orgName> and <address> elements
ISSN	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/idno[@type=’ISSN’]	@type value may be ISSN (generic) pISSN (printed version) or eISSN (electronic version)
Volume	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/imprint/biblScope[@type=’vol’]	
Issue	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/imprint/biblScope[@type=’issue’]	
First page	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/imprint/biblScope[@type=’fpage’]	
Last page	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/monogr/imprint/biblScope[@type=’lpage’]	
Type	/TEI/teiHeader/fileDesc/sourceDesc/biblStruct/@type	Possible values are: **article**, inproceeding, inbook, book, thesis, report
Subject headings	/TEI/teiHeader/profileDesc/textClass/keywords	When available, provided as a <list> of <item> for each keyword (defaults to what is provided in the Journal table)
Language	/TEI/teiHeader/profileDesc/langUsage/language/@ident	ISO 639-1 (defaults to 'en')
Embargo	/TEI/teiHeader/fileDesc/publicationStmt/availability	Note that the information is rarely provided. (defaults to what is provided in the Journal table)

Dealing with affiliations

Element <author>

It contains the information about the author, personal or corporate, of an article.

It groups three elements: the name of the author (<persName>), the affiliation (<affiliation>) and the email address.

Ex.:

```
<author>
    <persName>
       <forename>J.</forename>
       <surname>Kwela</surname>
    </persName>
    <affiliation>
       <orgName type="department">Institute of Experimental Physics</orgName>
       <orgName type="institution">University of Gdansk</orgName>
       <address>
         <addrLine>ul. Wita Stwosza 57</addrLine>
         <postCode>80-952</postCode>
         <settlement>Gdansk</settlement>
         <country key="PL">Poland</country>
       </address>
    </affiliation>
    <email>fizjk@univ.gda.pl</email>
</author>
```

Element <affiliation>

The <affiliation> component of <author> is intended to contain any potentially relevant information with regard to the author's academic situation: research group, laboratory, institution. It contains an informal description of the author's present or past affiliation with some organization, for example an employer or a sponsor.

It groups two elements: the name of the organization (<orgName>) and its postal address (<address>).

Ex.:

```
<affiliation>
      <orgName type="institution">Technische Universität Darmstadt</orgName>
      <orgName type="department">Institute of Materials Science</orgName>
      <address>
            <addrLine>Petersenstraße 23</addrLine>
            <postCode>D-64287</postCode>
            <settlement>Darmstadt</settlement>
            <country key="DE">Germany</country>
      </address>
</affiliation>
```

We have identified three types of organizations, which correspond to the three-tiered system of WoS.

- Institution: corresponds to the global structure that hosts the author (can be a university or an institute (e.g. MIT, INRIA)) – the largest scale of organization type.

- Department: corresponds to a specialized division of the institution mentioned above - intermediate structure of organization type (department, faculty, institute) if there is one.
- Laboratory: corresponds to the research team or group, which the author belongs to (e.g. Joint Research Laboratory Nanomaterials) – the smallest scale of organization type.

Element <address>

It contains the postal address of the organization, which the author is affiliated to.

Ex.:

```
<address>
      <addrLine>Chemin du Solarium</addrLine>
      <addrLine>BP20</addrLine>
      <postCode>33175</postCode>
      <settlement>Gradignan</settlement>
      <country key="FR">France</country>
</address>
```

Here are the different elements that can be used to describe the address.

- <addrLine> contains one line of a postal address. It can be used as many times as needed (in case of a multiple line address).
- <postCode> contains a numerical or alphanumeric code used as part of a postal address to simplify sorting or delivery of mail.
- <settlement> contains the name of a settlement such as a city, town, or village identified as a single geo-political or administrative unit.
- <region> contains the name of an administrative unit such as a state, province, or county, larger than a settlement, but smaller than a country.
- <country> contains the name of a geo-political unit, such as a nation, country, colony, or commonwealth, larger than or administratively superior to a region and smaller than a bloc. The key attribute must be used to identify the country, according to ISO 3166 a2.

Complex cases

Multiple institutions or multiple departments

a) This case typically corresponds to the affiliation of a joint laboratory.

Use the @key attribute to identify the different institutions (or departments), which the joint laboratory belongs to.

Example: Joint Research Laboratory Nanomaterials, which is a joint laboratory of the Technische Universität Darmstadt and the Karlsruhe Institute of Technology.

```
<affiliation>
  <orgName type="laboratory">Joint Research Laboratory Nanomaterials</orgName>
  <orgName type="institution" key="instit1">Technische Universität Darmstadt</orgName>
  <orgName type="institution" key="instit2">Karlsruhe Institute of Technology</orgName>
  <address>
      <addrLine>Petersenstrasse 23</addrLine>
      <postCode>D-64287</postCode>
      <settlement>Darmstadt</settlement>
      <country key="DE">Germany</country>
  </address>
</affiliation>
```

b) In the case of a laboratory described by several names (that is often the case for the UMRs of French CNRS), use the same procedure.

Example: GREMI, also named UMR 6606 (a joint laboratory of CNRS and Université d'Orléans)

```
<affiliation>
      <orgName type="laboratory" key="lab1">GREMI</orgName>
      <orgName type="laboratory" key="lab1">UMR 6606</orgName>
      <orgName type="institution" key="instit1">CNRS</orgName>
      <orgName type="institution" key="instit2">Université d'Orléans</orgName>
      <address>
            <addrLine>14, rue d'Issoudun</addrLine>
            <addrLine>BP 6744</addrLine>
            <postCode>45067</postCode>
            <settlement>Orléans cedex 2</settlement>
            <country key="FR">France</country>
      </address>
</affiliation>
```

c) In an even more complex case, where several departments belonging to different institutions can be found, a @corresp attribute is used to identify which superior structure each inferior unit belongs to. The @corresp attribute expresses a correspondence between the organisation and its parent structure, identified by the attribute *xml:id*.

Please note that when no @corresp attribute is used, it is implied that the laboratory belongs to the next superior structure (full distributivity).

Ex.1: Dipartimento di Fisica belongs to the Università di Trento, whilst Gruppo Collegato di Trento belongs to the INFN.

```
<affiliation>
      <orgName type="department" key="dep1" corresp="#org01">Dipartimento di
      Fisica</orgName>
      <orgName type="department" key="dep2" corresp="#org02">Gruppo Collegato
      di Trento</orgName>
      <orgName type="institution” key="instit1” xml:id="org01">Università di
      Trento</orgName>
      <orgName type="institution” key="instit2” xml:id="org02">Istituto
      Nazionale di Fisica Nucleare</orgName>
      <address>
      ...
      </address>
</affiliation>
```

Ex.2: IHCP is a department of the Joint Research Center, which is a department of the European Commission.

```
<affiliation>
      <orgName type="department" key="dep1" corresp="#org03">IHCP</orgName>
      <orgName type="department" key="dep2" xml:id="org03">Joint Research
Center</orgName>
      <orgName type="institution”>European Commission</orgName>
      <address>
      ...
      </address>
</affiliation>[7]
```

Ex.3: IGFAE is a joint research department of USC and Xunta de Galicia while Departamento de Fisica de Particulas is a department of USC.

```
<affiliation>
      <orgName type="department" key="dep1" corresp="#org04 #org05">IGFAE</orgName>
      <orgName type="department" key="dep2" corresp="#org04">Departamento de Fisica
      de Particulas</orgName>
      <orgName type="institution” key="instit1” xml:id="org04">USC</orgName>
      <orgName type="institution” key="instit2” xml:id="org05">Xunta de
      Galicia</orgName>
      <address>
      ...
      </address>
</affiliation>
```

Ex.4: Nancy I Poincaré is an institution belonging to a larger institution, Nancy Université.

```
<affiliation>
      <orgName type="department" key="dep1">Institut Jean Lamour</orgName>
      <orgName type="department" key="dep1">UMR 7198</orgName>
```

[7] In this case, it is implied that the Joint Research Center belongs to the only institution (European Commission). There is no need to use the @corresp attribute.

```
      <orgName type="institution” key="instit1” corresp="#org06">Nancy I
      Poincaré</orgName>
      <orgName type="institution” key="instit2” xml:id="org06">Nancy
      Université</orgName>
      <address>
      ...
      </address>
</affiliation>[8]
```

Several affiliations

The case is encountered when an author provides two affiliations (the current one and the one he wants to see appearing on the paper, which often corresponds to his past affiliation when he conducted the research that led to the article).

Use the attribute “current” to identify the current affiliation.

Example:

```
<author>
      <persName>
            <forename>B.</forename>
            <surname>Bastin</surname>
      </persName>
      <affiliation>
            <orgName type="laboratory">LPC Caen</orgName>
            <orgName type="institution" key="instit1">ENSICAEN</orgName>
            <orgName type="institution" key="instit2">Université de Caen
            Basse-Normandie</orgName>
            <orgName type="institution" key="instit3">CNRS/IN2P3</orgName>
            <address>
                  <postCode>14050</postCode>
                  <settlement>Caen</settlement>
                  <country key="FR">France</country>
            </address>
      </affiliation>
      <affiliation type="current">
            <orgName type="laboratory">GANIL</orgName>
            <orgName type="institution" key="instit4">CEA/DSM</orgName>
            <orgName type="institution" key="instit3">CNRS/IN2P3</orgName>
            <address>
                  <postCode>14076</postCode>
                  <settlement>Caen</settlement>
                  <country key="FR">France</country>
            </address>
      </affiliation>
```

[8] In this case, it is implied that the Institut Jean Lamour, also identified as UMR 7198, belongs to the next superior structure (instit1 - Nancy I Poincaré).

```
</author>
```

Several geographical addresses

For lack of encountered examples, this case will not be treated as such here.

Cases exist though and relate to laboratories, which are bilocalized and therefore have a double address for a single laboratory.

References:

- Holmes M., Romary L.: “Encoding models for scholarly literature” http://hal.archives-ouvertes.fr/hal-00390966/fr/
- PEER Final report on the provision of usage data and manuscript deposit procedures for publishers and repository managers (D2.2) http://www.peerproject.eu/fileadmin/media/reports/PEER__D2_2_20091028_v5.pdf
- TEI P5 Guidelines http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html
- http://www.inera.com/EML2002Rosenblum01.pdf