NEEO TECHNICAL GUIDELINES
DRAFT – Version 3.3

NEEO – WP5
Author: Benoit Pauwels

Date	Version	
15/10/2007	1.0	Initial document
30/10/2007	2.0	Integration of remarks made at WP5-TWG of 16/10/2007
14/11/2007	2.1	Some minor textual changes
19/2/2008	2.2	Added annex 2
15/4/2008	3.0	• Removed inconsistencies with MODS and DIDL application profile • Restructuring of document: o separate chapter on bibliographic metadata o object file metadata is now treated within chapter on "Object files" o chapter "Impact of NEEO specifications on IR" has been incorporated into other chapters of the document o "RePEc upload flag" moved under chapter on OAI • Added: o examples of valid DAI encodings o identifiers in a NEEO-compliant DID o modification date of the top-level DID item o a note on persistence of identifiers o a note explaingin that the RePEc OAI sets should be complete subsets of the NEEO OAI sets o recommended OAI identifier format o 'application/vnd.ms-powerpoint' as file format supported for full-text indexing o IR is responsible to set the top-level DID item modified-date correctly upon any relevant modification within the NEEO-DID o Paragraph on OCRised version of a PDF object file o Paragraph "Exposure of NEEO-DID through OAI" o First draft of annex 3 • Removed: o "Other local IR developments": no relation with NEEO application profile o "Persistent Identifiers": incorporated under "The NEEO Application Profile"
1/6/2008	3.1	• Version 0.5 of annex 3
18/8/2008	3.2	• Version 1.1 of annex 1 • Version 0.4 of annex 2 • Version 1.1 of annex 3 • Some minor editorial changes
26/2/2009	3.3	• Correction on p.18: unit="page" instead of unit="pages"

Outstanding issues:

- 3.2
  - Digital Rights Management (DRM): Creative Commons?
  - Sequence number of file within set of object files
- 4.10 and Annex 3: integrate the Repec story
- Annex 3: Repec ID of an author

Table of Contents

1 The NEEO application profile

Based on the findings of the Economists Online project (conducted by the NEREUS consortium between November 2005 and March 2006, funded by SURF), we have decided early on in this NEEO project to use the DIDL and MODS standards in order to express digital items, representing textual scientific publications. Please refer to the "WP5 Choosing for DIDL-MODS" document for a full report on the reasons for this choice.

This document describes the NEEO application profile, i.e. the way how to use the DIDL and MODS schemas in order to create a description of a scientific publication which guarantees maximum integration of these within the NEEO project and its end-user services. The NEEO application profile should therefore be understood as an aggregate of a DIDL and a MODS application profile, both of which are based on the corresponding application profiles developed by SURFshare (although NEEO introduces some extensions to these, as explained in the document underneath).

It is the desire of the NEEO project to develop and apply these profiles in synergy with other European initiatives in the digital library context, such as the DRIVER project, in order to reach a fully interoperable European network of institutional repositories and service providers.

This document is about textual publications only and not about datasets. This latter case is described in other guidelines that will be produced under the WP4 actions of the NEEO project.

In a first chapter we introduce the notion of a digital item and its representation as a DID (Digital Item Declaration), containing bibliographic metadata and (references to) the object files that it constitutes. In subsequent chapters we introduce the MODS application profile, the notions of object file metadata, the unique author identifier (called the DAI), the specifications that are to be followed for the implementation of the OAI-PMH protocol. The MODS and DIDL application profiles are fully explained in annexes 1 and 2 of these guidelines.

In a third annex the registration process of NEEO institutions and authors is described; which is fully based on an RDF/XML Schema, using the FOAF RDF vocabulary.

1.1 A digital item and its representations

An institutional repository is a software platform that permits researchers and academic staff to deposit their electronic publications and related digital material. In this process of deposit, the objects (files) of the electronic publication are electronically stored, together with additional information that describe (the contents of) these objects. The combination of one or more object files together with metadata is called a **digital item**. As we will see later the components of such a digital item can also be seen as digital items.

The content of any digital item (contained in an IR) can be semantically described through **bibliographic or descriptive metadata**, such as title, author(s), abstract, keywords, date of publication, specific identifiers, etc, and can contain zero or more **object files**:

- if the item is just a bibliographic reference for a resource, no object files are attached
- in the case of a complex work, an item can contain (for example) as many object files as there are chapters in the work
- a document can be made available in different formats (PDF, LaTeX, etc), each of these being a separate object file attached to the one digital item
- one object file can exist as different versions (postprint, publisher version, etc)

Each of these object files are described through so-called **object file metadata**, consisting of, for example, size and format of the file, restrictions to get access to the object, etc.

We can depict a digital item as in figure 1.

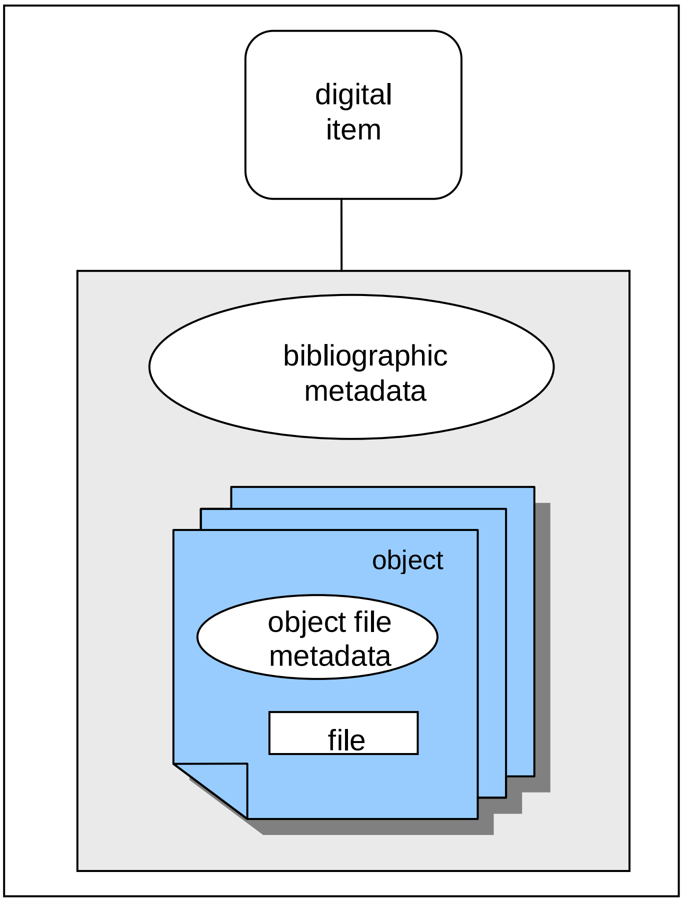


Figure 1: a digital item with its objects, bibliographic and object file metadata

A digital item can be represented in different ways. In a typical IR system this is done through some SQL database in combination with storage of the objects in a file system. One can also serialize digital items in XML, for example DIDL.

1.2 Representation of a digital item within an IR

Every IR software represents its digital items in a different way. As an example, within DSpace, the metadata (both bibliographical and object file) is stored in a set of PostgreSQL tables, and the objects reside in files on the file system. Consider the following article that sits in a DSpace system with three objects attached: the complete article in PDF format, chapter 1 in HTML, and chapter 2 in LaTeX:

> The geology and gold deposits of the Victorian gold province
> *Ore Geology Reviews*, *Volume 11, Issue 5*, *November 1996*, *Pages 255-302*
> G. Neil Phillips and Martin J. Hughes
> DOI: 10.1016/S0169-1368(96)00006-6

The internal DSpace record structure for the bibliographic metadata of this article would look like in figure 2. The DSpace internal ID for this item is ‘20’. The item was submitted by a person with id ‘5’. The bibliographic metadata of this item was last modified on 2004-12-29, and is stored according to the “qualified Dublin Core” data model. In a similar way object file metadata of the three attached objects is stored within the PostgreSQL database, like in figure 3.

item

item_id	submitter_id	last_modified
20	5	2004-12-29 15:55:55.85+01

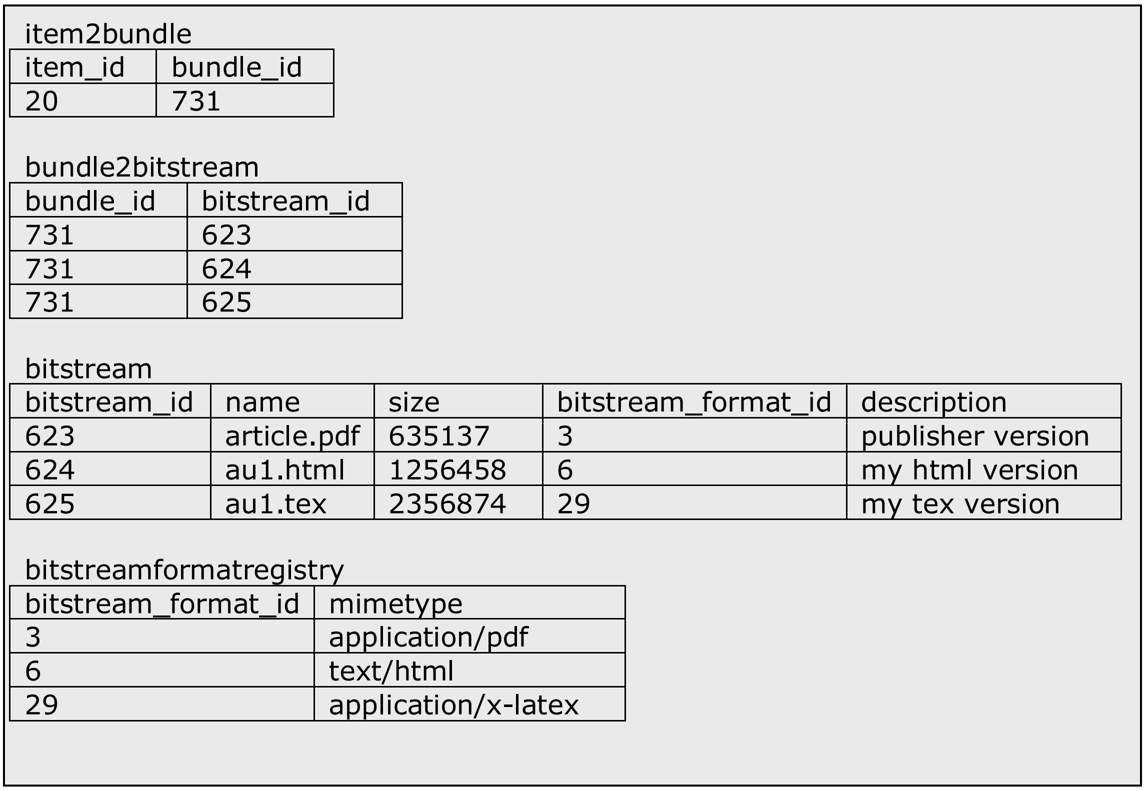
dctyperegistry

dc_type_id	Element	qualifier
1	contributor	
12	date	available
27	description	abstract
64	title	
18	identifier	citation
25	identifier	uri

dcvalue

item_id	dc_type_id	text_value
20	64	The geology and gold deposits of the Victorian gold province
20	1	Phillips, G. Neil
20	1	Hughes, Martin J.
20	12	11-1996
20	27	The Palaeozoic succession of Victoria represents a major world gold province with a total production of 2500 t of gold (i.e. 78 million oz). On a global scale, central Victoria …
20	18	Ore Geology Reviews, Volume 11, Issue 5, November 1996, Pages 255-302
20	25	10.1016/S0169-1368(96)00006-6

Figure 2: representation of bibliographic metadata in a DSpace system



item2bundle

item_id	bundle_id
20	731

bundle2bitstream

bundle_id	bitstream_id
731	623
731	624
731	625

bitstream

bitstream_id	name	size	bitstream_format_id	description
623	article.pdf	635137	3	publisher version
624	au1.html	1256458	6	my html version
625	au1.tex	2356874	29	my tex version

bitstreamformatregistry

bitstream_format_id	mimetype
3	application/pdf
6	text/html
29	application/x-latex

Figure 3: representation of object file metadata in a DSpace system

1.3 A digital item represented as a DIDL document

DIDL stands for "Digital Item Declaration Language", and permits for the representation of digital items in an XML format. With this language a digital item can in principle be represented in many ways (each of these being a so-called Digital Item Declaration (DID)). Within NEEO we have defined (through the NEEO application profile, see annex 2 of this document) a DID as an aggregate of three semantically different parts:

- the (bibliographic) metadata of the digital item
- the objects and their object file metadata; the objects are specified as links to them and are not stored as such within the DID
- a link to a so-called jump-off page, which is typically an HTML formatted intermediate page that is used for a human readable presentation of an item.

Graphically a DID can be depicted as follows:

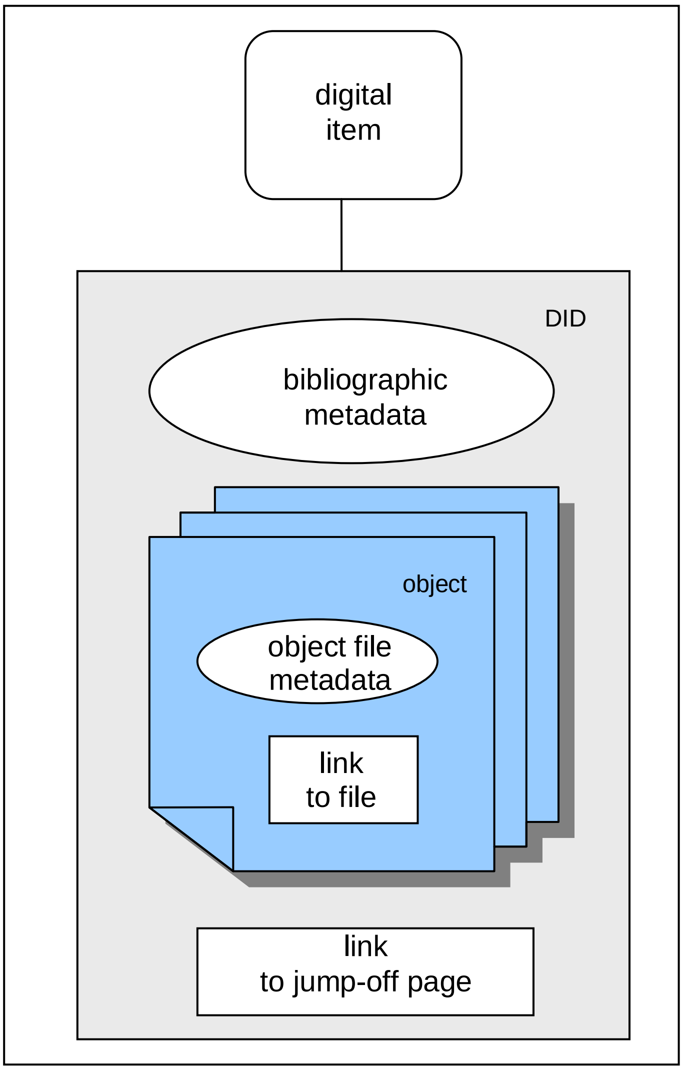


Figure 4: graphical representation of a digital item as a DIDL document

The above article would graphically look like this:

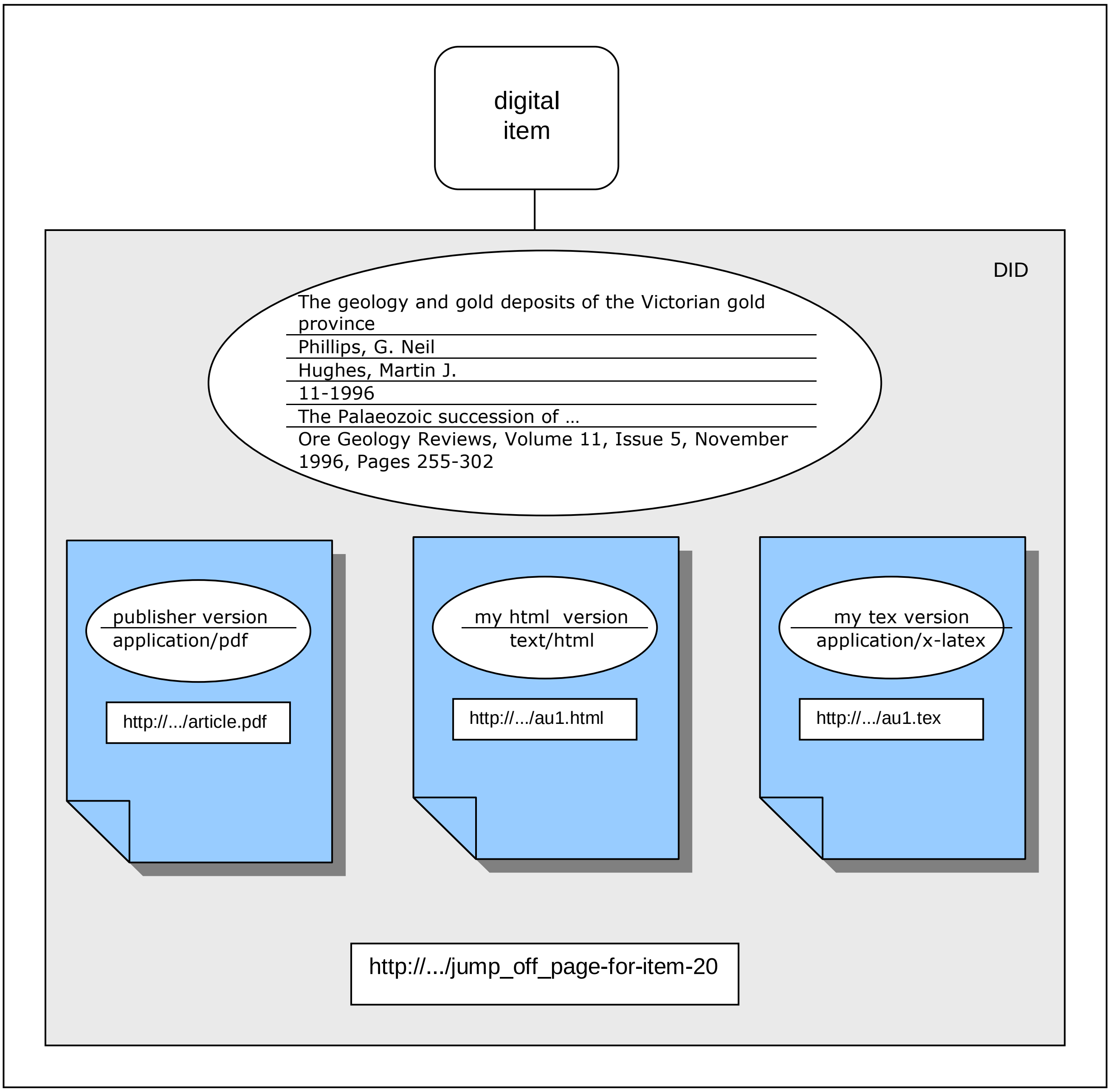


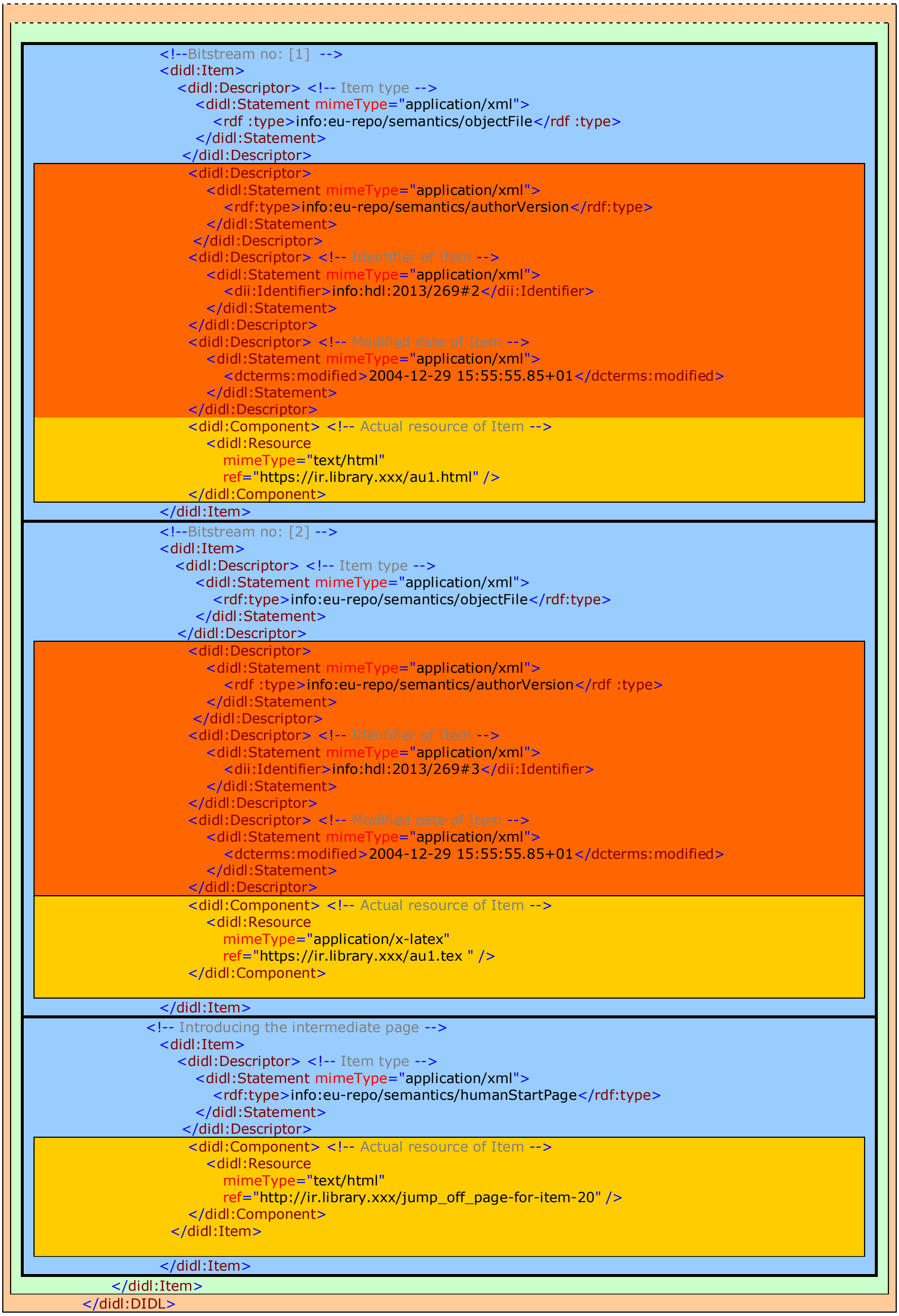
Figure 5: graphical representation of an 'article' digital item as a DIDL document

In its DIDL XML notation the above article would then look like this:

- *the following XML document conforms to the NEEO application profile as described in annexes 1 (use of MODS for the bibliographic metadata) and 2 (use of DIDL)*
- *identifiers are fictitious*

```
<didl:DIDL

  xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
  xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:dcterms="http://purl.org/dc/terms/"

  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="
    urn:mpeg:mpeg21:2002:02-DIDL-NS
      http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd

    urn:mpeg:mpeg21:2002:01-DII-NS
      http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/dii/dii.xsd"
>
    <!-- The Item is the autonomous compound entity that is a representation of a work-->
    <didl:Item>
        <didl:Descriptor>
          <didl:Statement mimeType="application/xml">
            <dii:Identifier>info:hdl:2013/269</dii:Identifier>
          </didl:Statement>
        </didl:Descriptor>
        <didl:Descriptor>
          <didl:Statement mimeType="application/xml">
            <dcterms:modified>2004-12-29 15:55:55.85+01</dcterms:modified>
          </didl:Statement>
        </didl:Descriptor>

        <!-- Introducing the area for metadata  -->
        <didl:Item>
          <didl:Descriptor> <!-- Item type  -->
            <didl:Statement mimeType="application/xml">
              <rdf:type>info:eu-repo/semantics/descriptiveMetadata</rdf:typeype>
            </didl:Statement>
          </didl:Descriptor>
          <didl:Descriptor>
            <didl:Statement mimeType="application/xml">
              <dii:Identifier>info:hdl:2013/269#mods</dii:Identifier>
            </didl:Statement>
          </didl:Descriptor>
          <didl:Descriptor>
            <didl:Statement mimeType="application/xml">
              <dcterms:modified>2004-12-29 15:55:55.85+01</dcterms:modified>
            </didl:Statement>
          </didl:Descriptor>
            <didl:Component> <!-- Actual resource of Item -->
              <didl:Resource mimeType="application/xml">
                <mods:mods
                  xmlns="http://www.w3.org/2001/XMLSchema"
                  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                  xmlns:mods="http://www.loc.gov/mods/v3"
                  xsi:schemaLocation="
                    http://www.loc.gov/mods/v3
                    http://www.loc.gov/standards/mods/v3/mods-3-2.xsd"
                >
                  <mods:titleInfo xml:lang="en‟>
                    <mods:title>The geology and gold deposits of the Victorian gold province
                     </mods:title>
                    <mods:nonSort>The</mods:nonSort>
                  </mods:titleInfo>
                  <mods:typeOfResource>text</mods:typeOfResource>
                  <mods:genre type="info:eu-repo/semantics/article‟ />
                  <mods:name type="personal‟ ID="_20n1‟>
                    <mods:namePart type="family‟>Phillips</mods:namePart>
                    <mods:namePart type="given‟>G. Neil</mods:namePart>
                    <mods:role>
                        <mods:roleTerm authority="marcrelator‟ type="code‟>aut</mods:roleTerm>
                    </mods:role>
                  </mods:name>
```

```
         <mods:name type=“personal‟ ID=“_20n2‟>
           <mods:namePart type=“family‟>Hughes</mods:namePart>
           <mods:namePart type=“given‟>Martin J.</mods:namePart>
           <mods:role>
              <mods:roleTerm authority=“marcrelator‟ type=“code‟>aut</mods:roleTerm>
           </mods:role>
         </mods:name>
         <mods:extension>
           <dai:daiList
             xmlns:dai="info:eu-repo/dai"
             xsi:schemaLocation="
               info:eu-repo/dai
               http://drcwww.uvt.nl/~place/SURFshare/dai-extension.xsd">
             <dai:identifier IDref="_20n1" authority="http://library.xxx/dai">
              1234567
             </dai:identifier>
             <dai:identifier IDref="_20n2" authority="http://library.xxx/dai">
              4523890
             </dai:identifier>
           </daiList>
         </mods:extension>
         <mods:abstract xml:lang=“en‟>
           The Palaeozoic succession of Victoria represents a major world gold province with a
           total production of 2500 t of gold (i.e. 78 million oz). On a global scale, central Victoria
           …
         </mods:abstract>
         <mods:originInfo>
           <mods:dateIssued>1996-11</mods:dateIssued>
         </mods:originInfo>
         <mods:language>
           <mods:languageTerm authority=“rfc3066‟ type=“code‟>en</mods:languageTerm>
         </mods:language>
         <mods:relatedItem type=“host‟>
           <mods:titleInfo><mods:title>Ore Geology Reviews</mods:title></mods:titleInfo>
           <mods:part>
             <mods:detail type=“volume‟><mods:number>11</mods:number></mods:detail>
             <mods:detail type=“issue‟><mods:number>5</mods:number></mods:detail>
             <mods:extent unit=“page‟>
               <mods:start>255</mods:start><mods:end>302</mods:end>
             </mods:extent>
           </mods:part>
         </mods:relatedItem>
         <mods:identifier type=“uri‟>info:doi/10.1016/S0169-1368(96)00006-6</mods:identifier>
       </mods:mods>
     </didl:Resource>
   </didl:Component>
</didl:Item>
<!-- Introducing the area for digital fulltext objects  -->
<!--Bitstream no: [0] -->
  <didl:Item>
    <didl:Descriptor> <!-- Item type -->
      <didl:Statement mimeType="application/xml">
        <rdf:type>info:eu-repo/semantics/objectFile</rdf:type>
      </didl:Statement>
    </didl:Descriptor>
     <didl:Descriptor>
       <didl:Statement mimeType="application/xml">
         <rdf:type>info:eu-repo/semantics/publishedVersion</rdf:type>
       </didl:Statement>
      </didl:Descriptor>
     <didl:Descriptor> <!-- Identifier of Item -->
       <didl:Statement mimeType="application/xml">
         <dii:Identifier>info:hdl:2013/269#1</dii:Identifier>
       </didl:Statement>
     </didl:Descriptor>
     <didl:Descriptor> <!-- Modified date of Item -->
       <didl:Statement mimeType="application/xml">
         <dcterms:modified>2004-12-29 15:55:55.85+01</dcterms:modified>
       </didl:Statement>
     </didl:Descriptor>
     <didl:Component> <!-- Actual resource of Item -->
       <didl:Resource
         mimeType="application/pdf"
         ref="https://ir.library.xxx/article.pdf" />
     </didl:Component>
  </didl:Item>
```



```
         <!--Bitstream no: [1]  -->
         <didl:Item>
           <didl:Descriptor> <!-- Item type -->
             <didl:Statement mimeType="application/xml">
               <rdf :type>info:eu-repo/semantics/objectFile</rdf :type>
             </didl:Statement>
           </didl:Descriptor>
            <didl:Descriptor>
              <didl:Statement mimeType="application/xml">
                <rdf:type>info:eu-repo/semantics/authorVersion</rdf:type>
              </didl:Statement>
             </didl:Descriptor>
            <didl:Descriptor> <!-- Identifier of Item -->
              <didl:Statement mimeType="application/xml">
                <dii:Identifier>info:hdl:2013/269#2</dii:Identifier>
              </didl:Statement>
            </didl:Descriptor>
            <didl:Descriptor> <!-- Modified date of Item -->
              <didl:Statement mimeType="application/xml">
                <dcterms:modified>2004-12-29 15:55:55.85+01</dcterms:modified>
              </didl:Statement>
            </didl:Descriptor>
            <didl:Component> <!-- Actual resource of Item -->
              <didl:Resource
                mimeType="text/html"
                ref="https://ir.library.xxx/au1.html" />
            </didl:Component>
         </didl:Item>
         <!--Bitstream no: [2] -->
         <didl:Item>
           <didl:Descriptor> <!-- Item type -->
             <didl:Statement mimeType="application/xml">
               <rdf:type>info:eu-repo/semantics/objectFile</rdf:type>
             </didl:Statement>
           </didl:Descriptor>
            <didl:Descriptor>
              <didl:Statement mimeType="application/xml">
                <rdf :type>info:eu-repo/semantics/authorVersion</rdf :type>
              </didl:Statement>
             </didl:Descriptor>
            <didl:Descriptor> <!-- Identifier of Item -->
              <didl:Statement mimeType="application/xml">
                <dii:Identifier>info:hdl:2013/269#3</dii:Identifier>
              </didl:Statement>
            </didl:Descriptor>
            <didl:Descriptor> <!-- Modified date of Item -->
              <didl:Statement mimeType="application/xml">
                <dcterms:modified>2004-12-29 15:55:55.85+01</dcterms:modified>
              </didl:Statement>
            </didl:Descriptor>
            <didl:Component> <!-- Actual resource of Item -->
              <didl:Resource
                mimeType="application/x-latex"
                ref="https://ir.library.xxx/au1.tex " />
            </didl:Component>

         </didl:Item>
       <!-- Introducing the intermediate page -->
         <didl:Item>
           <didl:Descriptor> <!-- Item type -->
             <didl:Statement mimeType="application/xml">
               <rdf:type>info:eu-repo/semantics/humanStartPage</rdf:type>
             </didl:Statement>
           </didl:Descriptor>
            <didl:Component> <!-- Actual resource of Item -->
              <didl:Resource
                mimeType="text/html"
                ref="http://ir.library.xxx/jump_off_page-for-item-20" />
            </didl:Component>
          </didl:Item>

         </didl:Item>
    </didl:Item>
</didl:DIDL>
```

Figure 6: Example of a DIDL document for an 'article' digital item

The above DIDL document can be structurally presented as follows (the same colours are used as in the full-blown XML above):

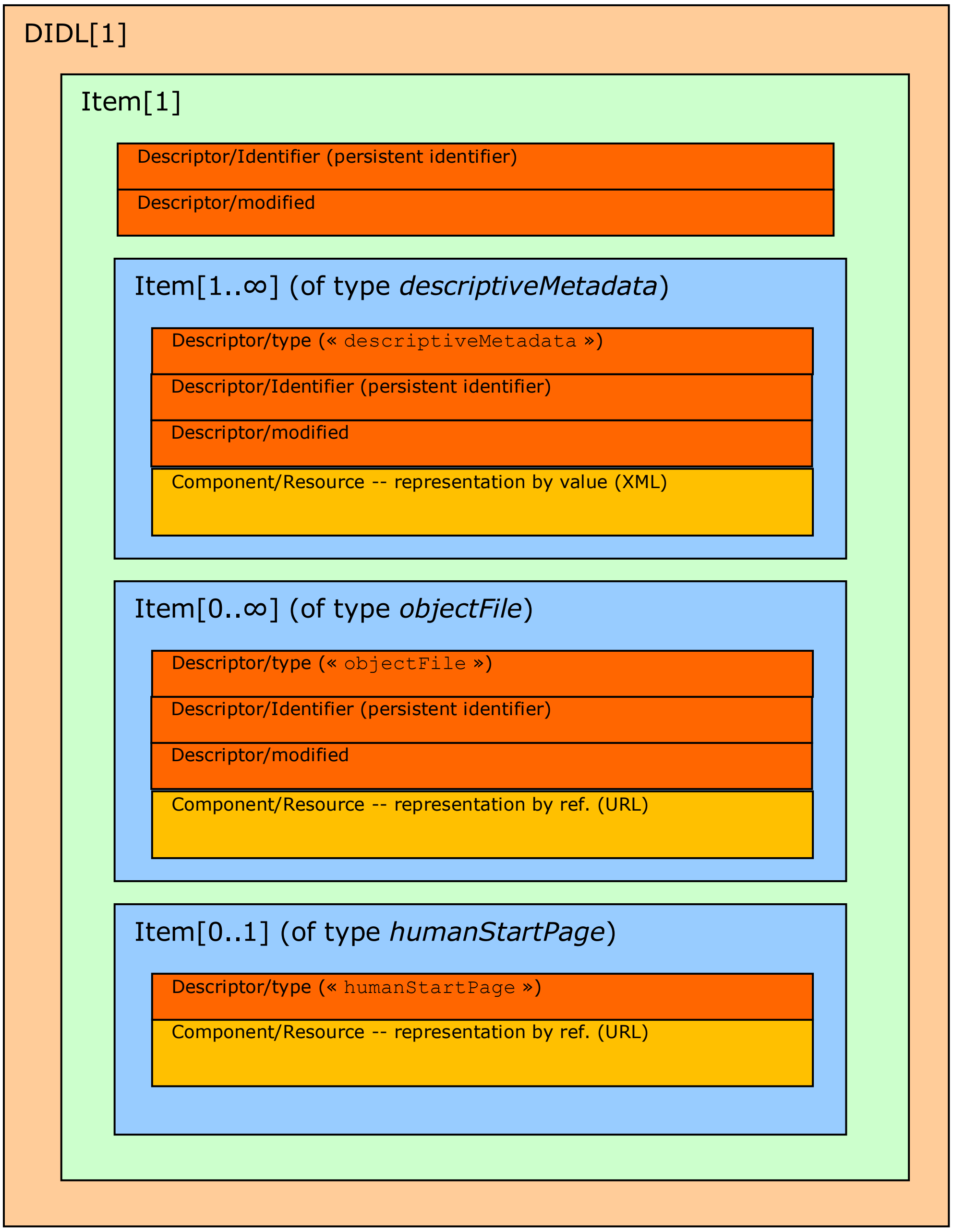

DIDL[1]
Item[1]
Descriptor/Identifier (persistent identifier)
Descriptor/modified
Item[1..∞] (of type *descriptiveMetadata*)
Descriptor/type (« descriptiveMetadata »)
Descriptor/Identifier (persistent identifier)
Descriptor/modified
Component/Resource -- representation by value (XML)
Item[0..∞] (of type *objectFile*)
Descriptor/type (« objectFile »)
Descriptor/Identifier (persistent identifier)
Descriptor/modified
Component/Resource -- representation by ref. (URL)
Item[0..1] (of type *humanStartPage*)
Descriptor/type (« humanStartPage »)
Component/Resource -- representation by ref. (URL)

Figure 7: Structure of a NEEO-compliant DIDL document

In the following we give a short introductory description of a NEEO-compliant DID. For a more detailed description, please refer to annex 2 of this document.

Every NEEO-compliant DID is composed of a maximum of three semantically different parts, called *Items* (the blue boxes), denoted by the *rdf:type* element, which can have 3 different values:

- `descriptiveMetadata`*:* this DIDL *item* holds a block of bibliographic metadata
- `objectFile`*:* this DIDL *item* holds a link to an object together with its object file metadata
- `humanStartPage`*:* this DIDL *item* holds a link to a jump-off page

descriptiveMetadata *Item*

- must contain an additional *descriptor* (red boxes) holding an identifier for the block of metadata, and can contain another (optional) *descriptor* which denotes the date at which the bibliographic metadata was last modified
- the bibliographic metadata is given by value in the DIDL document (i.e. the complete XML structure is included) in a *Component/Resource* element (orange box).
- the bibliographic metadata can be included multiple times in the DID (each in a separate *item* of *rdf:type* descriptiveMetadata), according to different data models (simple DC, QDC, MODS, etc). This permits IR managers to create ONE crosswalk that can comply with multiple application profiles. However, the NEEO application profile specifies that at least one of these bibliographic metadata parts must be in MODS according to the NEEO application profile for bibliographic metadata as specified in annex 1.

objectFile *Item*

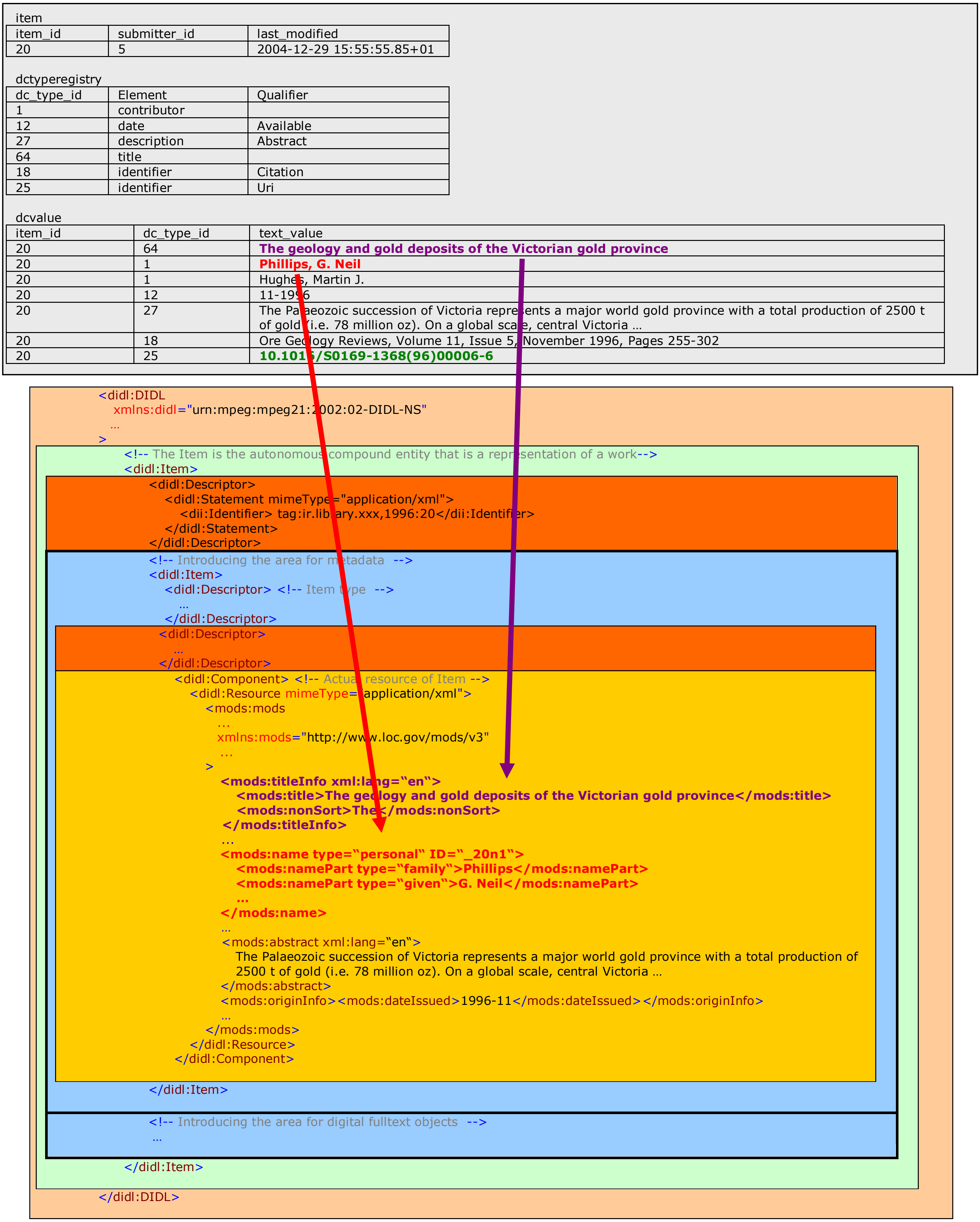
- must contain an additional *descriptor* (red boxes) holding an identifier for the object file, and can contain another (optional) *descriptor* which denotes the date at which the object file (or its metadata) was last modified
- the NEEO application profile specifies that each object of the digital item must be mentioned by reference, i.e. as a link to an object file (URL) in a *Component/Resource* element (orange box)
- each different object may be included in a separate *item* of *type* objectFile; if the objects have the same descriptors and the only difference is their media type, then the objects can be referenced in different Resource elements of the same Item/Component element
- "metadata-only digital items": one of the value-added services of the NEEO project is to build complete publication lists on a per-author basis, with links to the full-texts as far as allowed by copyright legislation. As such, we can expect that quite some digital items will be harvested by the NEEO gateway containing no reference at all to any object. This corresponds to DIDL documents that do not contain any objectFile *Item*s, but just one (or more) descriptiveMetadata *Item*(s) (and an optional humanStartPage *Item).*

humanStartPage *Item*

- does not contain any additional *descriptors*
- the NEEO application profile specifies that the jump-off page for the digital item must be mentioned by reference, i.e. as a link to an html page (URL) in a *Component/Resource* element (orange box)
- it is expected that the NEEO gateway will be able to construct such a page dynamically based on information contained in the other (metadata) parts of the DIDL document, and therefore the usage of such a jump-off page should become superfluous

1.4 Crosswalk between representations of a digital item

As explained above a digital item that is held in an institutional repository has its bibliographic and object file metadata structured in a different way than in a DIDL document. In order to transport the information about such a digital item to the NEEO gateway, one will need to convert from one representation (IR internal record structure) into the other (DIDL document that complies with the NEEO application profile). This process of converting is called a crosswalk. Graphically one could think of this process as in the following figure.



item

item_id	submitter_id	last_modified
20	5	2004-12-29 15:55:55.85+01

dctyperegistry

dc_type_id	Element	Qualifier
1	contributor	
12	date	Available
27	description	Abstract
64	title	
18	identifier	Citation
25	identifier	Uri

dcvalue

item_id	dc_type_id	text_value
20	64	**The geology and gold deposits of the Victorian gold province**
20	1	**Phillips, G. Neil**
20	1	Hughes, Martin J.
20	12	11-1996
20	27	The Palaeozoic succession of Victoria represents a major world gold province with a total production of 2500 t of gold (i.e. 78 million oz). On a global scale, central Victoria …
20	18	Ore Geology Reviews, Volume 11, Issue 5, November 1996, Pages 255-302
20	25	**10.1016/S0169-1368(96)00006-6**

```
<didl:DIDL
   xmlns:didl="urn:mpeg:mpeg21:2002:02-DIDL-NS"
   …
>
     <!-- The Item is the autonomous compound entity that is a representation of a work-->
     <didl:Item>
          <didl:Descriptor>
             <didl:Statement mimeType="application/xml">
                <dii:Identifier> tag:ir.library.xxx,1996:20</dii:Identifier>
             </didl:Statement>
          </didl:Descriptor>
          <!-- Introducing the area for metadata  -->
          <didl:Item>
             <didl:Descriptor> <!-- Item type  -->
                …
             </didl:Descriptor>
            <didl:Descriptor>
              …
            </didl:Descriptor>
               <didl:Component> <!-- Actual resource of Item -->
                  <didl:Resource mimeType="application/xml">
                     <mods:mods
                        …
                        xmlns:mods="http://www.loc.gov/mods/v3"
                        …
                     >
                        <mods:titleInfo xml:lang="en">
                           <mods:title>The geology and gold deposits of the Victorian gold province</mods:title>
                           <mods:nonSort>The</mods:nonSort>
                        </mods:titleInfo>
                        …
                        <mods:name type="personal" ID="_20n1">
                           <mods:namePart type="family">Phillips</mods:namePart>
                           <mods:namePart type="given">G. Neil</mods:namePart>
                           …
                        </mods:name>
                        …
                        <mods:abstract xml:lang="en">
                           The Palaeozoic succession of Victoria represents a major world gold province with a total production of
                           2500 t of gold (i.e. 78 million oz). On a global scale, central Victoria …
                        </mods:abstract>
                        <mods:originInfo><mods:dateIssued>1996-11</mods:dateIssued></mods:originInfo>
                        …
                     </mods:mods>
                  </didl:Resource>
               </didl:Component>
          </didl:Item>
          <!-- Introducing the area for digital fulltext objects  -->
           …
     </didl:Item>
</didl:DIDL>
```

Figure 8: Mapping of IR internal record structure into a NEEO-compliant DIDL document

1.5 Identifiers in a NEEO-DID

Several identifiers are used in a NEEO-compliant DID:

- The top-level, `descriptiveMetadata` and `objectFile` *Items* MUST have an identifier (expressed in a Descriptor element; see also annex 2 for further details)
- NEEO authors are identified through a DAI (expressed in the MODS extension element)
- Several other identifiers can be attached to an item of an IR, such as an identifier given to the item by the publisher (typically a DOI) or an identifier of the same work as stored in another IR. These identifiers are all part of the MODS metadata contained in a `descriptiveMetadata` *Item*.

In the following table we give an overview of the identifiers used. In the first column a semantic definition of the identifier is given. The second column holds the XPath expression of the location of the identifier within a NEEO-DID. In the last column we give an example of each of the identifiers for an item that is held in a DSpace IR. The DSpace IR software platform comes with the feature that permits to assign handles (according to http://www.handle.net) to items. In this specific example we assume a handle prefix of 2013; the DSpace item itself has the handle 2013/269.

Identifier	**Where**	**Example in DSpace**
identifier of version of publication as stored in repository; identifier is repository independent	/OAI-PMH/{GetRecord|ListRecords}/record/metadata/ DIDL/Item/Descriptor/Statement/dii:Identifier	info:hdl:2013/269
identifier of *descriptiveMetadata* part of publication as stored in repository; identifier is repository independent	/OAI-PMH/{GetRecord|ListRecords}/record/metadata/ DIDL/Item/Item[type="descriptiveMetadata"]/ Descriptor/Statement/dii:Identifier	info:hdl:2013/269#mods info:hdl:2013/269#dc
identifier of *objectFile* part of publication as stored in repository; identifier is repository independent	/OAI-PMH/{GetRecord|ListRecords}/record/metadata/ DIDL/Item/Item[type="objectFile"]/ Descriptor/Statement/dii:Identifier	info:hdl:2013/269#1 info:hdl:2013/269#2 info:hdl:2013/269#3 *the figure at the end is the 'internal sequence ID' for the object file*
identifier of an author of a publication	/OAI-PMH/{GetRecord|ListRecords}/record/metadata/ DIDL/Item/Item[type="descriptiveMetadata"]/ mods/extension/daiList/identifier	<dai:identifier authority= "http://library.xxx/dai"> 1234567 </dai:identifier>
identifier of publication as supplied by publisher	/OAI-PMH/{GetRecord|ListRecords}/record/metadata/ DIDL/Item/Item/mods/identifier	<mods:identifier> info:doi/10.1016/S0169 1368(96)00006-6 </mods:identifier> Other examples: info:pmid:... urn:isbn:...

For completeness (although this falls outside the scope of the DID), we also mention here the identifier of the OAI record, which transports the NEEO-DID. For further details on the format of the OAI identifier, see under "OAI Identifier", further down this document.

Identifier	Where	Example in DSpace
identifier of repository item as supplied by the repository	/OAI-PMH/request@identifier	oai:ir.library.xxx:2013/269

It is important that all these identifiers are persistent: once allocated to a DID item, author, OAI record, etc the identifier should remain unchanged. For obvious reasons: 2 examples:

- OAI identifier: if the same record is re-harvested by the NEEO gateway (for example as a consequence of a modification in the metadata, or the addition of a new object file), the NEEO gateway will update the record with the same OAI record identifier in its database. If the OAI record identifier of this record changes, this will not be possible, and the same record will therefore be wrongly present twice in the NEEO gateway.
- DID "top-level" item identifier: NEEO wants to enrich metadata of the item, by automatically attaching JEL classification codes, or by listing the references of an item. This will probably be done on a separate service, maintained by another instance than the NEEO Gateway. Other examples of enrichment could include translation services, usage metadata, etc. It is obvious that the merging of these different types of metadata of a publication in the interfaces of the NEEO gateway will only possible if the "top-level" DID item identifier remains unchanged.

1.6 Date modified of top level NEEO-DID item

This date should be the date at which the last relevant modification of any part of the NEEO-DID occurred. This could be a change in the bibliographic metadata, the addition of an object file, the change in the metadata of an object file, etc… It is up to the institutional repository to decide which modifications in the NEEO-DID should be defined as relevant. It is also their responsibility to propagate relevant modifications in the parts of the NEEO-DID document to the modified date of the top-level NEEO-DID item.

The OAI interface of the IR should use this top-level DID item modified-date to expose or not a NEEO-DID document. So, if a relevant modification occurs in a NEEO-DID document at date-time DT1 (and therefore the top-level DID item modified-date is set to DT1), then this NEEO-DID should be visible to the NEEO Gateway through the IR's OAI interface in response to an OAI ListRecords request stating "from=DT0", with DT0 <= DT1.

2 Bibliographic metadata

2.1 *Granularity*

The NEEO project has decided to opt for the MODS representation of the bibliographic metadata. The reasons for this choice have been explained in the "WP5 Choosing for DIDL-MODS" document. The MODS standard however permits for some flexibility in the way the different elements are fed. In order to ensure optimal interoperability between the data provider and the service provider, the NEEO project needs to set certain rules on how to use MODS for the description of scientific publications. A complete description of these rules can be found in the "NEEO application profile for bibliographic metadata" in annex 1.

An example: consider the article with DOI: `doi:10.1016/j.tranpol.2007.04.008`, available, for example, on the ScienceDirect platform at this URL: http://dx.doi.org/doi:10.1016/j.tranpol.2007.04.008 . Imagine that a post-print version of this article is available from a NEEO IR of an institution that houses Jan-Dirk Schmöcker (one of the authors of the article) and that this author has agreed to put his complete list of publications online through the NEEO Gateway.

- The author with the name '`Jan-Dirk Schmöcker`' should be encoded as follows according to the NEEO application profile, i.e. the last name of the person should go into the `<namePart>` element with the `type` attribute set to `"family"`. Similarly the first name should be mapped into the `<namePart type="given">` element. On top of this a unique identifier[1] for the author should be given (in order for the NEEO gateway to be able to build a publication list of all the works of Jan-Dirk Schmöcker): put an extra attribute `ID`[2] on the `<name type="personal">` element, and add an `<extension>` element as shown underneath:

```
<name type="personal" ID="n1">
    <namePart type="family">Schmöcker</namePart>
    <namePart type="given">Jan-Dirk</namePart>
    <role><roleTerm authority="marcrelator" type="code">aut</roleTerm></role>
</name>
<mods:extension>
    <daiList xmlns="info:eu-repo/dai"
              xsi:schemaLocation="
                        info:eu-repo/dai
                        http://drcwww.uvt.nl/~place/Surfshare/dai-extension.xsd">
        <identifier IDref="n1" authority="http://dai.thisuniversity.eu">1234567</identifier>
      </daiList>
</mods:extension>
```

Figure 9: Encoding of an author in MODS, according to the NEEO application profile

- The title '`The impact of the congestion charge on the retail business in London: An econometric analysis`' should be encoded like this (use the `titleInfo` element and the `title`, and `subTitle` subelements)

[1] called a Digital Author Identifier (DAI), for more details please refer to 2.2

[2] note that the IDs must be unique within the XML document. In our case the XML document is an OAI-PMH response that can contain several DIDL documents each with its own MODS metadata. One approach is to give each author within a MODS record a number, like n1, n2, ... and to prepend this with an underscore followed by the record id or item id in the repository.

```
<titleInfo>
  <title>The impact of the congestion charge on the retail business in London</title>
  <subTitle>An econometric analysis</subTitle>
</titleInfo>
```

Figure 10: Encoding of a title in MODS, according to the NEEO application profile

- The description of the journal in which this article has been published should be expressed in the following way, according to the NEEO application profile:

```
<originInfo>
  <dateIssued encoding=”iso8601”>2007</dateIssued>
</originInfo>
<relatedItem type=”host”>
  <titleInfo>
    <title>Transport Policy</title>
  </titleInfo>
  <part>
    <detail type=”volume”>
      <number>14</number>
    </detail>
    <detail type=”issue>
      <number>5</number>
    </detail>
    <extent unit=”page”>
      <start>433</start>
      <end>444</end>
    </extent>
  </part>
  <identifier type=”issn”>0967070X</identifier>
</relatedItem>
```

Figure 11: Encoding of the bibliographic citation (for an article) in MODS, according to the NEEO application profile

Similar rules are set for the other parts of the bibliographic metadata.

Following the guidelines in annex 1 will produce MODS structured metadata with a high level of granularity, which in turn permits the NEEO gateway to build qualitative added-value services on top of this metadata, such as the dynamic generation of publication lists per author, building of well-formed APA structured bibliographic references, etc.

2.2 Digital Author Identifier (DAI)

Building dynamic publication lists per author requires that these authors can be unambiguously identified. This is best done through a unique identifier that is assigned to each author of a work. Such an author identifier is called a DAI (Digital Author Identifier). There is no provision within ‘standard’ MODS for such an identifier. The NEEO project therefore defines this additional information in a MODS `<extension>` element, defined through a specific XML Schema: full details can be found in annex 1.

A DAI can be assigned to authors on a national level (like in the Netherlands where each author gets a unique identifier in the METIS system), or on an institutional level. The NEEO project doesn't want to get involved into assigning DAI's to authors: it is the sole responsibility of each IR to ensure that an author can be identified through a DAI and that each assigned DAI is unique within an IR.

Within the NEEO service, we have agreed to build complete publication lists for certain authors of our institutions: agreement needs to be obtained from at least 32 authors per institution for whom this publication list can be built. Each of these authors must be tagged with their DAI in the MODS metadata. Other authors may be identified through a DAI, but this is not an obligation.

2.2.1 Format of a DAI

The NEEO project does not impose any format for this DAI; every IR can deliver its DAI's in the format he wants as long as it validates against the XML Schema that can be found in annex 1. Within the NEEO service, all DAIs must be unique. This is accomplished by combining the DAI with its authority (value of the *authority* attribute of the *identifier* element) or by making the DAI a complete URI that is unique.

Some examples of valid encodings of a DAI:

```
<dai:identifier IDref="n1" authority="info:eu-repo/dai/nl">12456454</dai:identifier>

<dai:identifier IDref="n1" authority="staff.university.eu">19262</dai:identifier>

<dai:identifier IDref="n1" authority="http://staff.university.eu">19262</dai:identifier>

<dai:identifier IDref="n1">http://staff.university.eu/19262</dai:identifier>
```

Figure 12: Some examples of valid encodings of a DAI in MODS, according to the NEEO application profile

2.2.2 Persistence of a DAI

DAI's should be persistent identifiers: a change of DAI for an author could effectively result in incoherent results in certain services of the NEEO gateway. As an example publication lists in the NEEO gateway could become incomplete: part of the list would be allocated to DAI X, another part to DAI Y, both DAI's referring to the same author. Statistics on downloads of publications per author would also become incorrect.

If an institution needs to change the DAI's of its authors, for whatever reason, a complete re-harvest of the IR should be operated by the NEEO gateway, in order, for example, to get the publication lists right again. Errors in statistics would probably be irrecoverable. It is the responsibility of the IR administrator to advice the NEEO gateway administrator of the need of such a re-harvest operation.

The advice is clearly that DAI's shouldn't change, once they are assigned to authors.

2.2.3 Registration of a DAI in the NEEO gateway

The list of author DAIs (with some accompanying information on the authors, such as name, title, current affiliation, etc) needs to be delivered to the NEEO gateway administrator in a specific XML document. Refer to annex 3 for details on this.

2.2.4 Complementary author metadata

It would be a bad idea to incorporate author metadata (like email address, postal address, telephone number, link to biography, etc) in the MODS bibliographic metadata of a publication: this should contain only that information of an author that has some relation with the publication itself (i.e. the name of the author as it appears in the publication, a DAI, the role of the author in the publication (editor, main author, …)).

The current information of an author should be maintained in some system within the institution of the data provider (in some cases probably not maintained by the IR manager, but rather by some other administrative department), and that this information should be made available to the outside world through a web service. The NEEO gateway will use AJAX technology in its interfaces to present this information based on the DAI of the authors.

2.3 Bibliographic metadata structure in the IR

The NEEO requirement with respect to the quality of delivered metadata inevitably has its consequences on the metadata record structures (bibliographic and object file) for items in an IR. As an example, the DAI of the co-authors of a work need to be present within the bibliographic data structures of the IR (an alternative solution could be that this kind of information is dynamically generated by the OAI frontend of the IR upon harvesting by the NEEO gateway. But still this DAI needs to be available to the IR software, be it within its own data structures, or within a complimentary database which maintains information about the institution's authors).

Our DSpace example from above would therefore need to be described with a higher level of granularity in its metadata, if it wants to take part in the high-quality services of NEEO. One solution for this is to introduce the notion of subfields within the QDC fields of DSpace: each subfield denotes a separate information entity within the field. Example:

```
|aPhillips|jG. Neil|=24266
```

Each subfield (denoted by a '|' character followed by 1 other character) introduces a different kind of information:

```
|a    family name of the author
|j    first name of the author
|=    DAI of the author
```

The full DSpace example from above would then look as in the following figure.

This 'subfield' solution has been adopted by ULB: for each type of document (article, book chapter, working paper, etc) a complete set of fields and subfields has been defined, (mainly) based on the ISO 690[3] and ISO 690-2[4] standards. The latest version of the document describing this bibliographic metadata structure is available at

[3] ISO 690 : http://www.collectionscanada.ca/iso/tc46sc9/standard/690-1e.htm

http://www.bib.ulb.ac.be/RDIB/DISpace/DISpaceQDCfields.doc.

It is the sole responsibility of each NEEO data provider to make sure that a sufficiently high level of granularity is present within the metadata structures (both bibliographic and object file), so that a DIDL document can be delivered according to the NEEO application profile upon harvesting by the NEEO gateway. The solution adopted at ULB can be used as a guideline for some.

item

item_id	submitter_id	last_modified
20	5	2004-12-29 15:55:55.85+01

dctyperegistry

dc_type_id	Element	qualifier
1	contributor	
12	date	available
27	description	abstract
64	title	
18	identifier	citation
25	identifier	uri

dcvalue

item_id	dc_type_id	text_value
20	64	|aThe geology and gold deposits of the Victorian gold province
20	1	|aPhillips|jG. Neil|=24266
20	1	|aHughes|jMartin J.|=45238
20	12	|a11-1996
20	27	|aThe Palaeozoic succession of Victoria represents a major world gold province with a total production of 2500 t of gold (i.e. 78 million oz). On a global scale, central Victoria …
20	18	|aOre Geology Reviews|v11|i5|d11-1996|p255-302
20	25	|a10.1016/S0169-1368(96)00006-6

Figure 13: representation of bibliographic metadata in a DSpace system, with higher level of granularity

[4] ISO 690-2 : http://www.collectionscanada.ca/iso/tc46sc9/standard/690-2e.htm

3 Object files

3.1 File format

The added-values of the NEEO gateway include full-text searching and enrichment of the metadata of items based on the textual content of object files. This means that the object files need to be parsed and relevant information (like text, bibliographic references, JEL codes) needs to be recognized. This is not possible for all formats of object files. In the following we list the formats (with their corresponding IANA registered MIME Media Types) for which this parsing functionality can be guaranteed up to a certain level.

File format	**MIME Medium Type**
PDF	application/pdf
ODT	application/vnd.oasis.opendocument.text
TXT	text/plain
HTML	text/html
[La]TeX	application/x-latex
PostScript	application/postscript
MS-Word	application/msword
MS-PowerPoint	application/vnd.ms-powerpoint

This medium type of the object file is given through the `mimeType` attribute of the Component/Resource element within an `objectFile` *Item*.

```
<didl:Item> <!-- First Item for a File/Bitstream -->
    …
    <didl:Component>
      <didl:Resource
        mimeType="application/pdf"
        ref="http://my.server.nl/report.pdf"/>
    </didl:Component>
</didl:Item>
```

Figure 14: indicating the mime type of an object file in a NEEO-DID

3.2 Object file metadata

The objective of the NEEO search service is to be able to present a rich end user interface that contains relevant information about the object files of an IR item, in such a way that the user can easily decide (without additional clicks of the mouse), for example, which object files to download. This should be possible based on different criteria such as:

- an indication of the last modification date that points the user to the most recent version of an object file
- a general note about the content of an object file ("introduction", "chapter1", …)

- indication of version (postprint, publisher version, …)
- if no object files are present for a digital item, the user should be immediately informed about this
- access restrictions on an object file should be made clear to the user ("access forbidden", "publisher embargo", "open access", …)

We already explained above that each `objectFile` *Item* contains two Descriptor elements, denoting the identifier of the object file and a date of last modification. We also saw how to indicate the mime type of the object file.
In addition, the NEEO application profile defines the following object file metadata:

- a general description of (the semantic content of) the object file
- an indication of the version of the object file (postprint, publisher version, …)
- date at which the object file becomes available (e.g. after an embargo period set by the publisher)
- indication of accessibility of the object file
- deposit date of the object file

Each of these object file metadata elements is implemented as an additional *descriptor* of the `objectFile` *Item*. Please refer to annex 2 for more information.

```
<didl:Item> <!-- Second Item for a File/Bitstream -->
  <didl:Descriptor>
    <didl:Statement mimeType="application/xml">
      <rdf:type>
        info:eu-repo/semantics/objectFile
      </rdf:type>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!-- This Object Item has its own persistent ID -->
    <didl:Statement mimeType="application/xml">
      <dii:Identifier>urn:nbn:nl:ui:13-36724784</dii:Identifier>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!-- This Item has its own Modification date -->
    <didl:Statement mimeType="application/xml">
      <dcterms:modified>2006-12-20T10:29:12Z</dcterms:modified>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!-- a general description -->
    <didl:Statement mimeType="application/xml">
      <dc:description>publisher version</dc:description>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!—- accessibility of the object file -->
    <didl:Statement mimeType="application/xml">
      <rdf:type>
        info:eu-repo/semantics/openAccess
      </rdf:type>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!—- version of the object file -->
    <didl:Statement mimeType="application/xml">
      <rdf:type>
        info:eu-repo/semantics/publishedVersion
      </rdf:type>
    </didl:Statement>
  </didl:Descriptor>
  <didl:Descriptor> <!-- date of deposit -->
    <didl:Statement mimeType="application/xml">
      <dcterms:issued>2006-12-20T10:29:12Z</dcterms:issued>
    </didl:Statement>
  </didl:Descriptor>
```

```
    ...
    <didl:Component>
      <didl:Resource
        mimeType="application/pdf"
        ref="https://ir.library.xxx/article.pdf"/>
    </didl:Component>
  </didl:Item>
```

Figure 15: objectFile Item of a NEEO-DID with several object file metadata elements

3.3 OCR

The NEEO project wants to deliver a full-text indexing of the textual publications that it harvests from the IRs, and therefore wants to extract as much as possible text from the various file formats that NEEO data providers are submitting (as listed under 3.1).

In the case of the PDF file format, the file can contain the scanned/image or the OCRised/textual version of a publication. And yet, both versions can be denoted through the `application/pdf` mime type. It is the ambition of NEEO to do OCR of the scanned/image PDF files: NEEO will automatically determine whether an `application/pdf` document needs to be OCRised or not. NEEO data providers are therefore not obliged to OCRize their PDF publications themselves.

3.4 Metadata-only digital items

Metadata-only digital items will be harvested by the NEEO gateway.

It is not mandatory within the NEEO project for digital items that are contained within the IR to have object files attached: it is indeed possible that for certain publications the full-text cannot be found any more or that copyright legislation prevents a researcher from depositing the full-text in the IR.

3.5 Accessibility restrictions

The NEEO gateway will harvest digital items from an IR, whether or not the object files that are attached are openly accessible or not. NEEO of course encourages open access.

3.6 Object file metadata structure in the IR

This NEEO requirement with respect to the quality of delivered object file metadata again has its consequences on the metadata record structures for object files contained in the IR. In the DSpace example from above, for example, the object file metadata needs to be revisited, if it wants to comply with the NEEO application profile for object files. ULB has adopted the same 'subfield' solution as in the case of bibliographical metadata: in the following figure the same three object files are represented, however with additional metadata.

The following subfields have been defined:

```
|a    general description of object file
|d    deposit date of object file
|s    accessibility of object file (controlled vocabulary)
|t    date of end of embargo
```

```
|v    version of object file (controlled vocabulary)
```

Example: the first object file has the following metadata

```
|apublisher version|d1996-11|sULBINTERNET|t2000-01-01|vpublishedVersion
```

This is a publisher version (`|vpublishedVersion`) of the publication, which was deposited in the IR in November 1996 (`|d1996-11`). This object file is openly accessible for download from 1/1/2000 onwards (`|sULBINTERNET|t2000-01-01`). The general description for the file is 'publisher version' (`|apublisher version`).

A complete overview of object file metadata, as defined within the ULB DSpace system, can be found in the document http://www.bib.ulb.ac.be/RDIB/DISpace/DISpaceQDCfields.doc.

It is the sole responsibility of each NEEO data provider to make sure that a sufficiently high level of granularity is present within the object files metadata structure, so that a DIDL document can be delivered according to the NEEO application profile upon harvesting by the NEEO gateway. The solution adopted at ULB can be used as a guideline for some.

item2bundle

item_id	bundle_id
20	731

bundle2bitstream

bundle_id	Bitstream_id
731	623
731	624
731	625

bitstream

bitstream_id	name	size	bitstream_format_id	description
623	article.pdf	635137	3	|apublisher version|d1996-11|sULBINTERNET|t2000-01-01|vpublishedVersion
624	au1.html	1256458	6	|amy html version|d1994-12-14|sULBINTERNET|vauthorVersion
625	au2.tex	2356874	29	|amy tex version|d1994-12-14|sULBINTERNET|vauthorVersion

bitstreamformatregistry

bitstream_format_id	mimetype
3	application/pdf
6	text/html
29	application/x-latex

Figure 16: representation of object file metadata in a DSpace system, with higher level of granularity

4 OAI

The transportation of the DIDL documents from an IR to the NEEO gateway is done according to the OAI-PMH protocol, in its version 2.0. In the following we summarize specific guidelines for the implementation of an OAI-PMH interface on your IR.

4.1 OAI metadata crosswalk

All IR software platforms probably come with an implementation of the OAI-PMH 2.0 protocol. Within this OAI interface, every NEEO data provider needs to develop a metadata crosswalk that takes care of the conversion of the IR internal representation of a digital item into a DIDL document that complies with the specifications of the NEEO application profile (see also 1.4).

4.2 Identify response

Should contain information on the following:

- **Deleted records**

The NEEO gateway supports “deleted records”. The data provider needs to specify in the Identify response whether it supports this feature or not.

- **AdminEmail**

Every NEEO data provider must provide name and email of a person responsible for the correct functioning of the OAI interface of its IR. In case of malfunctioning of the harvesting process, this person will be asked to check XML validation errors in the logs generated on the NEEO gateway, and will be asked to correct bugs in the implementation of the crosswalk on its IR.

A typical OAI Identify response from a NEEO IR would then look like this:

```
<Identify>
   <repositoryName>DI-fusion</repositoryName>
   <baseURL>http://difusion.ulb.ac.be:8080/dspace-oai/request</baseURL>
   <protocolVersion>2.0</protocolVersion>
   <adminEmail>difusion-help@ulb.ac.be</adminEmail>
   <earliestDatestamp>2001-01-01T00:00:00Z</earliestDatestamp>
   <deletedRecord>persistent</deletedRecord>
   <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
   <description>
      …
   </description>
</Identify>
```

Figure 17: Example of an OAI Identify response, with indication of adminEmail and deletedRecord policy

4.3 OAI identifier

OAI sets can overlap, i.e. one item in the IR can belong to several OAI sets. The OAI protocol permits this. Since NEEO partners have the possibility to declare several OAI

sets for harvesting by the NEEO Gateway, we need to make sure that items that are transmitted as part of several OAI sets have the same OAI identifier in the different OAI sets, since otherwise the same OAI record would be present multiple times in the NEEO Gateway. Also, since the IR would be harvested (in part or completely) at frequent intervals, it is extremely important that identical OAI records are exposed under the same OAI identifier. This is a local responsibility: OAI identifiers should be unique and persistent, at least within the NEEO community, and worldwide.
It is decided that, in order to comply with these ideas above, the OAI identifiers used in NEEO should adhere to the "OAI identifier format" as described on http://www.openarchives.org/OAI/2.0/guidelines-oai-identifier.htm.

4.4 OAI set(s)

The concept of OAI sets gives the possibility to an IR administrator to divide its IR content into logical collections. A service provider, who needs to harvest records from such an IR, can specify the OAI set(s) it wants to harvest through the set parameter on the `listRecords` and `listIdentifiers` OAI verbs.

NEEO IRs can use this facility to limit harvesting to a part (or several parts) of their IR content. It is indeed very probable that an IR contains material outside of the economics topic, which therefore doesn't need to be incorporated into the NEEO gateway.

Every IR manager should decide how to create these OAI sets in function of his IR software platform. This could be accomplished, for example, through the periodical execution of some script that creates/updates a specific OAI set with item identifiers that correspond to:

- publications of "NEEO authors", who are each identified through a DAI
- and/or publications that are tagged with a specific classification code or subject topic
- and/or all authors that are known within the institution as producing scientific output in the field of economics
- etc

It is important to consider that NEEO wants to harvest not only metadata of publications of "NEEO authors", but also the current output in the economics topic: it is therefore not sufficient to create OAI set(s) solely based on the DAIs of the "NEEO authors".

If an IR contains only digital items that all need to be harvested by the NEEO gateway, then no OAI sets need to be created on the IR: the NEEO gateway will simply harvest the complete IR.

NEEO does not do any filtering: all records that are harvested from an OAI repository or from sets thereof will be ingested into the NEEO gateway, and therefore be presented through the various NEEO services.

The NEEO project does not impose any specifications for the setSpec and names of these OAI sets. These need to be declared to the NEEO gateway administrator. Please refer to annex 3 for the procedure on how to do this.

4.5 metadataPrefix naming

The NEEO gateway harvests records from the data providers using a combination of the `listRecords` and `getRecord` OAI requests. These 2 OAI verbs use the `metadataPrefix` parameter to obtain the metadata of the digital item in a specific representation.

All NEEO data providers **must** expose their metadata through the OAI interface under the metadataPrefix "didl".

4.6 Resumption token lifespan

All NEEO data providers should respect a reasonable time for a resumption token to be kept alive: at least 24 hours.

4.7 Harvest batch size

Every NEEO data provider should deliver OAI records in batches of a reasonable size up to a maximum of 200.

4.8 Exposure of NEEO-DID through OAI

The OAI interface of the IR should use the top-level DID item modified-date to expose or not a NEEO-DID document. So, if a relevant modification occurs in a NEEO-DID document at date-time `DT1` (and therefore the top-level DID item modified-date is set to `DT1`), then this NEEO-DID should be visible to the NEEO Gateway through the IR's OAI interface in response to an OAI `ListRecords` request stating "`from=DT0`", with `DT0 <= DT1`.

4.9 Frequency of harvesting

Two modes of harvesting are possible:

- Bulk: the complete IR (or the declared OAI sets thereof) is harvested by the NEEO gateway
- Incremental: the NEEO gateway harvests creations and modifications of digital items in the IR at periodical intervals

Frequency of (bulk and/or incremental) harvesting needs to be agreed upon with the EO support desk at Tilburg University (economistsonline@uvt.nl).

4.10 "RePEc upload" flag

Another value-added service that the NEEO project wants to implement is the automated upload of metadata of publications contained in the NEEO gateway to RePEc. This is an optional service for which every individual IR can opt or not: it is indeed possible that an IR is already uploading metadata to RePEc about (some or all of) its publications. For such an IR it would be undesirable that metadata about these publications would be send a second time to RePEc through NEEO.

Therefore every publication for which the IR wants NEEO to send the metadata to RePEc should be flagged as such. This needs to be done by 'putting' the publication in a specific OAI set of which the setSpec needs to be declared with the NEEO administrator (the way to do this declaration is explained in annex 3). In an OAI record this OAI set would then show up like in the following example:

```
<record>
   <header>
      <identifier>oai:difusion.ulb.ac.be:2013/474</identifier>
      <datestamp>2005-02-22T01:00:17Z</datestamp>
      <setSpec>hdl_2013_95</setSpec>
   </header>
   <metadata>
      <didl:DIDL …>
         …
      </didl:DIDL>
   </metadata>
</record>
```

With “hdl_2013_95” being the setSpec of the OAI set of publications for which metadata needs to be forwarded to RePEc

Figure 18: “RePEc upload” flag is indicated through a setSpec in the header of an OAI record

Important note

The “RePEc OAI sets” should be complete subsets of the “NEEO OAI sets”. The NEEO gateway will only harvest the “NEEO OAI sets”, and NOT the “RePEc OAI sets”: those items in the NEEO OAI sets that also fall in the RePEc OAI sets will be forwarded to RePEc. Every NEEO partner remains of course free to put items in the RePEc OAI sets without putting them in the NEEO OAI sets. Only, these items will not be forwarded to RePEc through NEEO.

5 XML validation of ingested OAI records

All ingested OAI records are validated against the XML Schemas that are used within the record (OAI-PMH, MPEG-21/DIDL, MODS, DAI, etc).
Ingested records that fail to validate are refused in the NEEO gateway. The administrator of the original IR is advised by the EO support desk (economistsonline@uvt.nl) of this shortcoming through an email message.

6 Annexes

6.1 Annex 1: Use of MODS for institutional repositories

http://drcwww.uvt.nl/~place/neeo/Use%20of%20MODS%20for%20institutional%20repositories-version%201.1.doc

6.2 Annex 2: MPEG21 DIDL Document Specifications for repositories

http://drcwww.uvt.nl/~place/neeo/didl%20application%20profile.doc

6.3 Annex 3: Registration of NEEO IR and authors

- Overview of registration process and detailed description of the XML Admin file
  http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5 Technical guidelines - Annex3.pdf

- XML Admin template file
  http://homepages.ulb.ac.be/~bpauwels/NEEO/WP5/WP5 Technical guidelines - Annex3Template.rdf