Content and Data Architecture, Operations
Elsevier
Radarweg 29
1043 NX Amsterdam
Netherlands

Phone: +31 20 312 2316
Email: j.migchielsen@elsevier.com

The ConSyn schemas

ConSyn_Schemas — version 0.11, 17 July 2014, by Jos Migchielsen

version 0.1	15 July 2010	First draft
version 0.2	29 July 2010	Second draft
version 0.3	27 August 2010	Third draft
version 0.4	22 August 2012	Updated to ConSyn schemas v0.6
version 0.5	15 May 2013	Updated to ConSyn schemas v0.6, element `prism:embargoDate` added
version 0.6	2 September 2013	Added support for open access
version 0.7	4 December 2013	Minor changes in description of embargo and open access elements
version 0.8	24 April 2014	Updated to ConSyn schemas v0.7, element `oa:windowType` added
version 0.9	3 June 2014	Updated to ConSyn schemas v0.8, element `cp:oa-last-updated-timestamp` added
version 0.10	4 July 2014	Updated to ConSyn schemas v0.9, support for stage S280 added
version 0.11	17 July 2014	Updated to ConSyn schemas v0.10, support for 5.4.0 DTDs added

Contents

1 Introduction

ConSyn is an application that contains XML content from Elsevier's Electronic Warehouse in an XML database. It allows customers to select and identify interesting journal articles, book chapters, etc. and order or download the identified content.

The XML files ConSyn delivers are *enhanced*, i.e. they contain metadata in the form of RDF. Additionally they can contain "raw text". Unlike XML deliveries from the Electronic Warehouse, the XML documents adhere to a W3C schema instead of a DTD. There is one set of schemas for ConSyn input and ConSyn output files.

This document describes the schemas and their elements.

Note: The described schemas are not the final versions. They will be revisited in the near future.

2 ConSyn schemas

The content ConSyn delivers are journal articles structured according to the Journal Article DTD version 4.5.2 (converted articles), 5.0.1, 5.0.2, 5.1.0, 5.2.0 or 5.4.0, and book hub files and book chapters structured according to the Elsevier Book DTD version 5.1.0, 5.1.1, 5.2.0, 5.2.1, 5.3.0, 5.3.1 or 5.4.0. Because the JA DTDs $5.x$ and the Book DTDs are backwards compatible, *one* equivalent schema is used for every type. This results in the following three schemas:

```
art452.xsd
art540.xsd
book540.xsd
```

Elements that these DTDs have in common are placed in the Common Element Pool (CEP). The DTDs mentioned here use various versions of the CEP. Because of backwards compatibility again one equivalent schema is used, or rather one set of schemas. The CEP elements are subdivided as follows: the ”core”, the elements for structured affiliations, the elements for structured bibliographic elements, the MathML elements, with Elsevier modifications, the CALS tables elements and the Extended CALS elements. These are declared in different DTDs. (See the *Tag by Tag* for more details.) The equivalent schemas are the following:

```
common140.ent.xsd
common140-soextblx.xsd
sa.xsd
sb.xsd
mathml2-mod-ES.xsd
mathml2-qname-1.mod.xsd
soextblx.xsd
tb.xsd
xlink.xsd
xml.xsd
```

The enhancement is in the form of RDF. It contains elements from the *PRISM 2.0* (http://www.prismstandard.org) and *Dublin Core* (http://www.dublincore.org) standards. Open access information is stored in elements belonging to the Elsevier Open Access namespace. Copyright and license information is stored in elements belonging to the Elsevier Copyright namespace. The corresponding schemas are:

```
rdf.xsd
prism.xsd
dct.xsd
oa.xsd
cp.xsd
```

The enhancement and the item XML are delivered as one file, and the schema that describes that file is used for both ConSyn input and output. The *document* schema describes this file. It contains elements with information to create the RDF on output (*document properties*) as well

Namespace prefix	Namespace
doc	http://www.elsevier.com/xml/document/schema
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
dp	http://www.elsevier.com/xml/common/doc-properties/schema
cps	http://www.elsevier.com/xml/common/consyn-properties/schema
oa	http://vtw.elsevier.com/data/ns/properties/OpenAccess-1/
cp	http://vtw.elsevier.com/data/ns/properties/Copyright-1/
cja	http://www.elsevier.com/xml/cja/schema
ja	http://www.elsevier.com/xml/ja/schema
bk	http://www.elsevier.com/xml/bk/schema
prism	http://prismstandard.org/namespaces/basic/2.0/
dct	http://purl.org/dc/terms/
ce	http://www.elsevier.com/xml/common/schema
cals	http://www.elsevier.com/xml/common/cals/schema
tb	http://www.elsevier.com/xml/common/table/schema
mml	http://www.w3.org/1998/Math/MathML
sa	http://www.elsevier.com/xml/common/struct-aff/schema
sb	http://www.elsevier.com/xml/common/struct-bib/schema
xml	http://www.w3.org/XML/1998/namespace
xlink	http://www.w3.org/1999/xlink

Table 1: Namespaces and namespace prefixes used in the ConSyn schemas.

as elements used for ConSyn processing (*ConSyn properties*). Therefore, the following three schemas complete the set of 21 ConSyn schemas:

```
document.xsd
dp.xsd
cps.xsd
```

See Section 4 for more details on the schemas.

3 Namespaces

Table 1 lists the various namespaces used by the ConSyn schemas with the namespace prefixes that Elsevier uses.

4 Details

The top element of the main ConSyn schema, `doc:document` contains the following four subelements:

- `rdf:RDF`

- dp:document-properties
- cp:consyn-properties
- one of the (converted) journal article or book top elements

Each is described in more detail in the following subsections.

4.1 rdf:RDF

The rdf:RDF element is only present in ConSyn-out files. It is based on the Elsevier web-PDF specifications v6.1, specifically the part on the XMP metadata.

rdf:RDF has one mandatory subelement rdf:Description. It contains the elements described below, in that order. The subelements are optional except where mentioned. rdf:Description has a mandatory attribute about which contains the DOI of the item (and is empty when that DOI is not present).

dct:format is *mandatory* and contains application/xml.

dct:title contains the title (and label and subtitle, if present) of the item.

dct:creator contains an author of the item. There is one dct:creator for every author.

```
<dct:creator>Richard G Trohman</dct:creator>
<dct:creator>Michael H Kim</dct:creator>
<dct:creator>Sergio L Pinski</dct:creator>
```

dct:subject contains all the keywords of the item that are in the default keyword class. The element contains one subelement rdf:Bag which contains a subelement rdf:li for every keyword.

```
<dct:subject xmlns:dct="http://purl.org/dc/terms/">
  <rdf:Bag>
    <rdf:li>aortic stenosis</rdf:li>
    <rdf:li>transapical</rdf:li>
    <rdf:li>transcatheter aortic valve implantation</rdf:li>
    <rdf:li>transfemoral</rdf:li>
  </rdf:Bag>
</dct:subject>
```

dct:description contains the citation data of the item: the journal or book title, volume number or edition number, year, page numbers, and the DOI.

```
<dct:description>The Lancet 364 (2004) 1701-1719.
  doi:10.1016/S0140-6736(04)17358-3</dct:description>
```

prism:aggregationType is *mandatory* and indicates the type of publication. It contains

journal or book.

prism:publicationName is *mandatory* and contains the title of the journal or book. In case of book series, the series title is used.

prism:edition contains the edition number of the book. It is omitted in case there is no edition information or if it is a first edition.

```
<prism:edition>Second Edition</prism:edition>
```

prism:copyright is optional and contains the copyright line of the item. It is not present when there is an *empty* copyright notice.

```
<prism:copyright>Copyright © 2004 Elsevier Ltd All rights reserved.</prism:copyright>
```

dct:publisher is *mandatory* and contains the name of the publisher of the publication.

prism:issn contains the ISSN of the publication.

prism:isbn contains the ISBN of the publication. In case of journals both ISSN and ISBN can be present. In case of books there can be more than one ISBN.

prism:volume contains the volume number(s) in which the item is published.

prism:number contains the issue number(s) in which the item is published.

prism:issueIdentifier contains the supplement code. For instance, S1 for Supplement 1, PA for Part A.

prism:coverDisplayDate contains the date that is displayed on the cover.

prism:coverDate contains the cover date in the format yyyy-mm-dd. It is only present when there is *one* date which can be transformed to this format.

prism:issueName contains a book volume or journal special issue title.

prism:pageRange contains the page range(s) of the item.

prism:startingPage contains the start page of the item (in case of multiple page ranges this is the first start page).

prism:endingPage contains the end page of the item (in case of multiple page ranges this is the last end page).

```
<prism:pageRange>13-37, 111<prism:pageRange>
<prism:startingPage>13</prism:startingPage>
<prism:endingPage>111</prism:endingPage>
```

prism:doi contains the DOI of the item.

prism:url contains the DOI of the item, but in URL form based on the DOI.

dct:identifier contains the DOI.

```
<prism:doi>10.1016/S0140-6736(04)17358-3</prism:doi>
<prism:url>http://dx.doi.org/10.1016/S0140-6736(04)17358-3</prism:url>
<dct:identifier>doi:10.1016/S0140-6736(04)17358-3</dct:identifier>
```

oa:openAccessInformation contains the open access information in subelements oa:openAccessStatus, oa:openAccessEffective, oa:sponsor (subelements oa:sponsorType and optional oa:sponsorName), (optional) oa:userLicense and (optional) oa:windowType. The open access status, the sponsor type, the user license and the window type are URIs.

cp:licenseLine contains the license line of the item.

4.2 dp:document-properties

The dp:document-properties element contains information about an item that is not present in the item itself. It is used in the delivery of ConSyn-out files. It is present in ConSyn-in files. In ConSyn-out files it can only contain subelements dp:raw-text and dp:version-number.

dp:document-properties contains the elements described below, in that order. The subelements are optional except for dp:aggregation-type.

dp:raw-text contains the so-called raw text of the item, essentially a text dump of the item. It is copied over to dp:document-properties/dp:raw-text in the output.

dp:aggregation-type indicates the type of publication. It contains journal, book, book series or mrw. It is used to fill rdf:RDF/prism:aggregationType. (The values book series and mrw are translated to book.)

prism:embargoDate contains the embargo date/time of the item. The type is xs:dateTime, i.e. a date/time formatted as specified in ISO 8601. Times are in seconds. The date/time should be in UTC, indicated by suffix "Z". For example, 2013-05-15T09:39:01Z. The element is filled as part of the ConSyn ingestion process; the information is taken from the VTW Properties Store. The element can be updated as a result of an hourly poll of the VTW Properties Store.

dp:version-number contains the version number of the item, e.g. S300.4. This can be an S280 version number.

dp:issue-hub-pii contains the PII of the issue hub file.

dp:book-hub-pii contains the PII of the book hub file.

dp:title contains the title (and label and subtitle, if present) of the item. It is copied over to

rdf:RDF/dct:title.

dp:issue-name contains a book volume or journal special issue title. It is copied over to rdf:RDF/prism:issueName.

dp:volume-issue-number contains the volume number(s), the issue number(s) and the issue's supplement code in the subelements dp:vol-first, dp:vol-last, dp:iss-first, dp:iss-last and dp:suppl. The information is used for elements rdf:RDF/prism:volume, rdf:RDF/prism:number and rdf:RDF/prism:issue Identifier.

dp:year-first contains the year of publication, or the first year of publication in case the issue or book is published in more than one year.

dp:year-last contains the last year of publication in case the issue or book is published in more than one year. It is omitted when there is only one year.

dp:item-pages contains the page-range(s) of the item. This information is used for rdf: RDF/prism:pageRange, rdf:RDF/prism:startingPage and rdf:RDF/prism: endingPage.

```
<dp:item-pages>
  <ce:pages>
    <ce:first-page>13</ce:first-page>
    <ce:last-page>37</ce:last-page>
  </ce:pages>
  <ce:pages>
    <ce:first-page>111</ce:first-page>
  </ce:pages>
</dp:item-pages>
```

ce:edition contains the edition number of the book the item is published in. It is copied over to rdf:RDF/prism:edition. The element is omitted if there is no edition information or if it is the first edition.

ce:copyright-line contains the copyright line of the item. It is copied over to rdf:RDF/ prism:copyright.

dp:publisher contains the Publisher of the journal or book. It is copied over to rdf:RDF/ dct:publisher.

ce:issn contains the ISSN of the publication. It is copied over to rdf:RDF/prism:issn.

ce:isbn contains the ISBN of the publication. It is copied over to rdf:RDF/prism:isbn. In case of books there can be more than one ISBN.

dp:cover-display-date contains the date that is displayed on the cover. It is copied over to rdf:RDF/prism:coverDisplayDate.

dp:cover-date contains the cover date in the format yyyy-mm-dd. It is copied over to rdf:RDF/prism:coverDate. It is omitted when there is a date range and when the date cannot be transformed to the format yyyy-mm-dd.

dp:item-sequence-nr contains the sequence number of the item in the journal or book.

dp:openAccessInformation contains the open access information in subelements oa:openAccessStatus, oa:openAccessEffective, oa:sponsor (subelements oa:sponsorType and optional oa:sponsorName), (optional) oa:userLicense and (optional) oa:windowType. The open access status, the sponsor type, the user license and the window type are URIs and are taken from lists of values. The elements are filled as part of the ConSyn ingestion process; the information is taken from the VTW Properties Store. The elements can be updated as a result of an hourly poll of the VTW Properties Store. The information is used for rdf:RDF/oa:openAccessInformation.

dp:copyrightInformation contains the copyright notice and the license line of the item in optional subelements cp:copyrightNotice and cp:licenseLine. The elements are filled as part of the ConSyn ingestion process; the information is taken from the VTW Properties Store. The elements can be updated as a result of an hourly poll of the VTW Properties Store. The information is used for elements rdf:RDF/prism:copyright and rdf:RDF/cp:licenseLine, respectively.

4.3 cps:consyn-properties

The element cps:consyn-properties contains information that is only used for ConSyn processing. Once a ConSyn-in file is processed it is stored in ConSyn *with* this element added. It is not present in ConSyn-in and ConSyn-out files.

cps:consyn-properties contains the elements below, in that order. All elements are optional.

cps:creation-timestamp contains the timestamp of the creation of the item in ConSyn.

cps:last-updated-timestamp contains the timestamp of the last update of the item in ConSyn.

cps:oa-last-updated-timestamp contains the timestamp of the last update of the item's open access information in ConSyn.

cps:loader-version is an internal ConSyn parameter indicating which version of the ConSyn ingestion is used.

cps:stage contains the stage of the item, e.g. S100.

cps:version-number contains the version number of the item, e.g. S300.2. This can be an S280 version number. It is copied over to dp:version-number in the output.

cps:md5 contains the md5 checksum of the file *without* the cps:consyn-properties

element.

4.4 ja:article,...

Element doc:document will contain as last subelement one of the following:

```
cja:converted-article
ja:article
ja:simple-article
ja:book-review
ja:exam
bk:book
bk:chapter
bk:simple-chapter
bk:examination
bk:fb-non-chapter
bk:glossary
bk:index
bk:introduction
bk:bibliography
```

Although above the schemas were called equivalent to the DTDs, this is not completely true. Two examples:

– In CEP 1.1.6 the attribute view was added to element ce:abstract (with default value all). That CEP was only used by the Book DTD $5.3.x$ and not by any Journal Article DTD. Hence when a ConSyn-out file containing a journal article is viewed *together* with the ConSyn schemas the ce:abstract element will suddenly contain a view attribute with value all. Without the ConSyn schemas the attribute is of course not present.

– In the Journal Article DTD element ce:link has an attribute locator of type ENTITY. ce:link is used to link to images, MMCs, etc. These entities are declared at the top of the item. The ConSyn-out file does not contain those declarations and therefore the type of the attribute was changed to string. The attribute now contains the filename. In practice the filename and the locator are the same and an effort is being made to ensure this is always the case.

Note also that all the journal and book elements now carry a namespace prefix.