National Library of Sweden

Date: 2015-03-25
Version: 2.2
Ref:

Legal deposit of electronic documents by means of RSS feeds

National Library of Sweden

Date: 2015-03-25
Version: 2.2
Ref:

Change history

Version 2.2

- An optional element *dcterms:identifier* with attribute *@xsi:type* has been added to enable registration of **identifier type**, improving the description of the resource and facilitating deduplication.
- The elements *dcterms:isPartOf* and *dcterms:references* may also have attribute *@xsi:type* to indicate identifier type of reference.

Version 2.1

- An optional element *dcterms:isPartOf* has been added to enable registration of **host publication** at article level, under <*item*>.

Version 2.0

- Mandatory elements **no longer** need to be registered in a given order. This is an adaptation to XML Schema 1.1, permitting random order between mandatory elements.
- XPath-expressions for rss have been corrected and are now given without prefix, since RSS has no namespace of its own.
- S201 *media:content* is now mandatory, if applicable. This concerns in particular images, audio files, video clips associated with a resource (item).
- S201 **Premium object (file) in a** *media:group* **-** *//media:group/media:content/@isDefault* in version 1.3 is canceled. We now fetch all objects in a *media:group*, irrespective of content.
- An element *dcterms:references* has been added to enable designation of a media file delivered through another channel, e.g via *ftp*.
- F308 **Holder of rights pertaining to object (file) / contributor** has been split into two separate elements: **(F308)** Creator of /**contributor to object –** *media:credit*, and **F308 Holder of rights pertaining to object** – *media:copyright*.

- A new section **5. Validation** has been added, with references to xml-schemas currently in force.
- Information about whether elements are *repeatable* (R) has been added for each element / attribute. (R) means here that an element can have multiple occurrences *under the same< item>*.

Version 1.3

- Mandatory elements must now come first under each item in a given sequence (adaptation to the corresponding xml-schema *rss_kb.xsd* now published).
- The element R105 <*title*> is now mandatory, while R106 <*description*> is optional.
- The element R107 <*dcterms:accessRights*> now gets one of two possible values: '*gratis*' or '*restricted*'.
- The elements R118 *media:keywords* and R118 *media:category* have been split into two tables each (R118 and F311); one is for the description of the resource (*item*) as a whole, the other table is for constituent files (objects), described as *media:content*.
- Support for retrieval via *https* is now implemented by KB.

Version 1.2

- Under the heading "Publishing": Namespace addresses for MediaRSS and Dublin Core is now correct.

Version 1.1

- The element *link* is now mandatory
- New element *dcterms:isFormatOf* (optional) added

Editorial change:

- References (ID-no and headings) to the document "Introduktion till metadata i leveranser av elektroniska dokument till KB" (in Swedish) at the top of most tables.

1. Introduction

This document describes how legal deposit suppliers of electronic documents can deliver material by means of RSS feeds.

2. Publishing

A supplier wishing to deliver via feeds publishes an RSS feed at a fixed URL. The frequency of retrieval is settled in an agreement with Kungl. biblioteket (KB) – The National Library of Sweden. The RSS feed needs to fulfil the following conditions:

- The format must be RSS 2.0 with the addition of some elements from MediaRSS 1.5.0 (http://search.yahoo.com/mrss/) and Dublin Core DCMI Metadata Terms (http://purl.org/dc/terms/).

  For MediaRSS the best information about the format is presently found here: http://www.rssboard.org/media-rss/. Please observe, however, that this is still only a draft document, as stated in the editor's note, while the link to the official specification at yahoo.com is currently inactive.

  For Dublin Core, please note that KB uses the more comprehensive namespace for DCMI Metadata Terms, *not* the more restricted Dublin Core Metadata Element Set (http://purl.org/dc/elements/1.1/), encompassing only 15 elements, being only a small subset of DCMI Metadata Terms. Some of the Dublin Core elements in this specification are exclusively part of the DCMI Metadata Terms. This concerns for example *accessRights* and *isFormatOf* in the format tables below (section 6).

  The prefix *dc* or *dct* is sometimes also used for the namespace of the more comprehensive DCMI Metadata Terms (http://purl.org/dc/terms/). This is accepted, as long as it is consistently used exclusively for this namespace. For the sake of clarity, however, we recommend the use of the prefix *dcterms* for all elements in Dublin Core.

Format	Commonly used prefix	Namespace URI	Documentation
RSS 2.0	-	-	http://www.rssboard.org/rss-specification
MediaRSS 1.5.0	*media*	http://search.yahoo.com/mrss/	http://www.rssboard.org/media-rss
DCMI Metadata Terms	*dcterms*	http://purl.org/dc/terms/	http://dublincore.org/documents/dcmi-terms/

Summary of recommended formats, prefixes and namespaces.

- Each item in the RSS feed represents an electronic document that is to be delivered. Metadata about the document is expressed in the formats RSS, MediaRSS and *dcterms* for each item. The files that make up the electronic document that is represented by an item are designated using the elements <link> from RSS [and], <media:content> from MediaRSS and dcterms:format. These designations contain information about where the files are published (URL) and their logical format (MIME-type). The files designated in this way should be accessible for retrieval by KB, at the URLs given. All URLs should be of type *http* or *https*.
- *Items* in the feeds must be sorted by the value in the element <pubDate> in descending order, i.e. with the latest published item first.
- The element <pubDate> must have as its value the time when the latest version of the document was published. If a new version of the document is published, the value of <pubDate> must be updated and the *item* put in the correct sequence further up in the feed.
- All *items* that are published or have new versions published during a time period corresponding to the interval of retrieval (settled in the agreement between KB and the supplier) must be shown in the feed.
- Please observe that even if a metadata element is optional according to this specification, this does not mean that the corresponding file(s) or object(s) described by this metadata element are exempt from legal deposit.

3. Limitations

Delivery by means of feeds has certain limitations:

- Delivery of files with DRM-protection is not supported.

4. Security

It is anticipated that suppliers may wish to deliver material that is not meant to be publically accessible and thus needs to be protected from unauthorized retrieval by a third party. This can be solved using one of the following methods:

- IP-filter – the supplier configures its network infrastructure so as to allow retrieval only from certain IP-addresses, used by KB only.
- Log in with Basic Authentication – the supplier protects the resources by username/password to be used by KB.

Support for retrieval via *https* is now implemented by KB.

5. Validation

In order to control that RSS-feeds to be delivered fulfil the requirements in this specification, KB provides xml-schemas that the supplier / publisher can use for validation in advance. These revised xml-schemas are based XML Schema version 1.1 and are available for download here: http://www.kb.se/namespace/rss/index.html .

For local validation to work properly, all these six different schemas (*1. rss-kbse.xsd, 2. dcterms_kbse.xsd, 3. media_kbse.xsd, 4. content_kbse.xsd, 5. dc.xsd, 6. dcmitype.xsd*) must be downloade to the same file folder. No. 1. is the main schema against which validation is then performed. In the validation engine or software used for local validation, select XML Schema **version 1.1** as default.

6. Format

This is a selection of elements from RSS, MediaRSS and *dcterms*, for which KB specifies certain requirements in order to make delivery via feeds work properly. Mandatory elements are marked with * and have a Yes in the Mandatory field. Repeatable elements have a (R) after their name. This means the element can occur several times *under the same <item>*.

Other parts of the given standards (for RSS MediaRSS and *dcterms*) may also be used in a feed and, if supplied, some of those will also be used by KB.

R101	***Identifier**
/rss/channel/item/guid	
Mandatory	Yes
Format	Optional
Unique and permanent identifier of the resource. Will be used for feedback to the supplier and to identify new versions of the same resource.	

R101a	**Identifier with type value (R)**
/rss/channel/item/dcterms:identifier[@xsi:type]	
Mandatory	No, but desirable for purpose of deduplication.
Format element	doi, ean, hdl, isan, isbn, ismn, isrc, issue-number, matrix-number, upc, uri, urn
Format @attribute	**Controlled values**: “dcterms:doi”,”dcterms:ean”, ”dcterms:hdl”, ”dcterms:isan”, ”dcterms:isbn” etc.
Unique and permanent identifier of the resource. Can be the same as in R101 *item/guid*. But multiple identifiers of different types are possible here (repeatable element). Type value given by means of attribute @xsi:type requires a reference to the namespace for xsi, either in the root-element <rss> or in the element where @xsi:type occurs. The type value must have the same prefix as is used for namespace http://purl.org/dc/terms/ in the root. **Example:** <rss xmlns:**dcterms**=http://purl.org/dc/terms/ version="2.0"> <**dcterms**:identifier xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:type="**dcterms**:isbn">9783452679123</dcterms:identifier>	

R102	***Internet address**
/rss/channel/item/link	
Mandatory	Yes
Format	URL
Internet address (URL) of the electronic resource at the time of publishing, that is where the resource was first made publicly available. May be used for feedback to the supplier and to identify new versions of the same resource. This URL will be used to retrieve one of the files that make up the document. The same information content that is designated by <link> may also be delivered in another format, without ads, navigation etc. This other format will then be given by an element <*media:content*>, having as a child-element <*dcterms:isFormatOf*> with the same URL value as <link>. (See below F302 and following element).	

R103	***Publishing date**
/rss/channel/item/pubDate	
Mandatory	Yes
Format	RFC822
Time of publishing. If the resource is republished in a new version, this date must be updated. All items in a channel should be sorted by this field in descending order, with newest item first.	

R104	***Publisher**
/rss/channel/item/dcterms:publisher	
Mandatory	Yes
Format	http://id.kb.se/organisations/SE[organisation number]+ [-suffix]
Unique identifier of institution (publisher) that made the resource available on the internet. Swedish *organisation number* is here written *without hyphen*. If several publishers (e.g. newspapers) share the same organisation number with a common supplier (distributor, owner publishing house), in agreement with KB a suffix *must* be added for reference to the particular publisher (newspaper) that published the item (article). The suffix must be preceded by a hyphen and hold minimum two characters in the range [A-Z,a-z,0-9]. **Example**: The Swedish online newspaper *Dalademokraten*, is delivered and owned by Mittmedia AB, with organization no. 556004-1815, which is shared by several other newspapers in the same publishing house. For all items published by *Dalademokraten* the reference to publisher, with the agreed suffix is then: <item> <dcterms:publisher>http://id.kb.se/organisations/SE**5560041815-DD** </dcterms:publisher> </item>	

R105	***Title**
/rss/channel/item/title	
Mandatory	Yes
Format	Text
Name of resource.	

R106	**Description**
/rss/channel/item/description	
Mandatory	No
Format	Text
Short description of content, e.g. introduction, abstract or summary of resource.	

R107	***Accessibility at the time of publishing**
/rss/channel/item/dcterms:accessRights	
Mandatory	Yes
Format	**Controlled values (either of two)**: *gratis* | *restricted*
Indicate if the resource is freely accessible (’*gratis*’) at the time of publication. In other case, if there are particular conditions for access to the resource (such as payment, subscription, encryption), select ‘*restricted’*.	

R108	**License**
/rss/channel/item/dcterms:license	
Mandatory	No
Format	URI, Text
Can be used to state a license encompassing all constituent parts of the document. This could be for example Creative Commons.	

R112	**Host publication (R)**
/rss/channel/item/dcterms:isPartOf[@xsi:type]	
Mandatory	No
Format element	doi, ean, hdl, isan, isbn, ismn, isrc, issn, issue-number, matrix-number, upc, uri, urn
Format @attribute	**Controlled values**: “dcterms:doi”,”dcterms:ean”, ”dcterms:hdl”, ”dcterms:isan”, ”dcterms:isbn”, “dcterms:issn” etc.
Used at item-level for reference to a host publication, e.g. an online journal in which an article is placed, in those cases where different host-publications occur in the same feed under different items. If an item has been published in several different host publications the element can be repeated. For further formatting instructions, see **example** under **R101a Identifier with type value.** Otherwise, as a default, the host publication is taken from *channel/link*.	

R115	**Statement of responsibility (Creator) (R)**
/rss/channel/item/dcterms:creator	
Mandatory	No
Format	**For names of persons:** Family name, Given name (role)
Name(s) of person(s) or institution(s) that are responsible for the intellectual or artistic content, for example author, photographer, composer.	

R115	**Statement of responsibility (Contributor) (R)**
/rss/channel/item/dcterms:contributor	
Mandatory	No
Format	**For names of persons:** Family name, Given name (role)
Name(s) of person(s) or institution(s) that contributed to the intellectual or artistic content of the resource, for example illustrator, translator.	

F303	***File format (for object designated by *link*)**
/rss/channel/item/dcterms:format	
Mandatory	Yes
Format	MIME
MIME-type of the file designated by the <**link**>-element	

S201	***Constituent objects (files) that make up the resource (R)**
/rss/channel/item/(media:group/)media:content	
Mandatory	Yes, if applicable
Format	-
Reference to constituent part of a resource (item). Particularly relevant for audio, video files, images belonging to the resource (with @type =”audio/…”, ”video/…”image/…” ; see further F303). Can be grouped in *media:group*. Also used for reference to a file with the same content as the resource designated by *item/link* above, but in another format, e.g. stripped from commercial advertisements, navigation etc. (with, for example, @type=”text/html”; see also R102, F302).	

F302	***Internet address of constituent object (R) /** **Retrieval address for alternative format**
/rss/channel/item/(media:group/)media:content/@url	
Mandatory	Yes, if *media:content* is provided
Format	URL
Internet address (URL) to constituent object (file) belonging to the electronic resource at the time of publication. Also used as URL for retrieval of alternative format of resource (see S201 above).	

-	**Indicator of alternative format**
/rss/channel/item/(media:group/)media:content/dcterms:isFormatOf	
Mandatory	No
Format	URL (same value as in <link>)
Same URL as in <link>. The presence of this element indicates that the parent element *media:content* describes the same content as in <link>, but in another format. (See above R102, S201 and F302).	

F303	***File format (for object in *media:content*)**
/rss/channel/item/(media:group/)media:content/@type	
Mandatory	Yes, if media:content is provided
Format	MIME
MIME-type (http://www.freeformatter.com/mime-types-list.html) of the file designated by this media:content. Composed of two parts a slash in between. The first part refers to a media type (e.g. “video”, “audio” or “image”), while the second part, after the slash designates the file format. **Examples**: *video/x-flv, audio/mpeg, image/jpeg.* To refer to the same resource as in <link> in an alternative format (without commercials etc.) e.g. @type=”text/html” can be used here. (See above 201).	

[S201]	***Object (media files) belonging to resource, but delivered as legal deposit separately (e.g. via FTP) (R)**
/rss/channel/item/dcterms:references[@xsi:type]	
Obligatoriskt	*Yes, if applicable
Format element	doi, ean, hdl, isan, isbn, ismn, isrc, issn, issue-number, matrix-number, upc, uri, urn
Format @attribute	**Controlled values**: “dcterms:doi”,”dcterms:ean”, ”dcterms:hdl”, ”dcterms:isan”, ”dcterms:isbn” etc.
Alternative for media files *not* included in *media:content*. For further formatting instructions, see **example** under **R101a Identifier with type value.**	

F305	**Checksum (R)**
/rss/channel/item/(media:group/)media:content/media:hash	
Mandatory	No
Format	MD5
An MD5 checksum for the designated file.	

F307	**License for a constituent object (file) (R)**
/rss/channel/item/(media:group/)media:content/media:license	
Mandatory	No
Format	URI
Can be used to indicate a license that is valid for a certain part of the document, for example Creative Commons.	

R118	**Keywords (subject) for *item* (R)**
/rss/channel/item/media:keywords	
Mandatory	No
Format	Comma separated list of keywords
Keywords relevant for the resource (*item*) as a whole.	

F311	**Keywords (subject) for constituent file (object) (R)**	
/rss/channel/item/(media:group/)media:content/media:keywords		
Mandatory	No	
Format	Comma separated list of keywords	
Keywords relevant for a constituent file (object) described as *media:content*.		

(F308)	**Creator of / contributor to object (file) (R)**	
/rss/channel/item/(media:group/)media:content/media:credit		
Mandatory	No	
Attribute	role	Role according to scheme
	scheme	urn:ebu | urn:yvs
Format	**For names of persons**: Family name, Given name (role)	
Name and role for the person (or institution) that has contributed to the artistic or intellectual content of an object (file) belonging to a resource.		

F308	**Holder of rights pertaining to object (file) (R)**
/rss/channel/item/(media:group/)media:content/media:copyright	
Mandatory	No
Attribute	url
Format	**For names of persons**: Family name, Given name (role)
Name and role for the person (or institution) that has contributed to the artistic or intellectual content of the resource.	

R118	**Category (subject) of *item* (R)**	
/rss/channel/item/media:category		
Mandatory	No	
Attribute	scheme	URI
	label	Category name in clear text
Format	Value according to selected scheme	
Can be used for categorizing the resource (*item*) as a whole.		

F311	**Category (subject) of constituent file (object) (R)**	
/rss/channel/item/(media:group/)media:content/media:category		
Mandatory	No	
Attribute	scheme	URI
	label	Category name in clear text
Format	Value according to selected scheme.	
Used for a category pertaining to a single file (an object) belonging to the resource.		