Journal of Computer Science 8 (3): 425-430, 2012
ISSN 1549-3636

Application of Syndication to the Management of Bibliographic Catalogs

Manuel Blazquez Ochando and Juan-Antonio Martinez-Comeche
Department of Library and Information Science,
Faculty of Library and Information Sciences, University Complutense,
Santísima Trinidad 37, 28010 Madrid Spain

Abstract: Problem statement: The process of transmission of bibliographic records between libraries is a complex task, usually treated by the Z39.50 protocol. **Approach:** The objective of this research is to propose an alternative method to simplify this process, using the techniques of content syndication. **Results:** The computer program compares the feasibility of using different formats (ATOM, RSS1.0, RSS2.0 and MARC-XML) to convey and share library catalogs of various sizes (up to 1 millon records) between libraries. Tests have shown that smaller collections of 25,000 records, the time insertion/import catalogs is less than 1 min. **Conclusion/Recommendations:** The analysis suggests that syndication is a useful technique for the transmission and retrieval of bibliographic information.

Key words: Content syndication, library catalogs, bibliographic management, automation of information centers, MARC-XML

INTRODUCTION

Content syndication refers to a technique for transmitting information in XML (Brickley and Guha, 2004) through channels or sources that can be updated and shared with any client on the web.

Dave Winer (2001) is one of the pioneers of this technique. It was applied initially to the mass media (New York Times, 2009). Now it has been extended to various areas, including the academic field, specializing in textual and audio-visual content.

Applications in Librarianship consist of channels for literature alert or with general information (ANU, 2010). It has been also used for the dissemination of articles and scientific journals (Abadal *et al*., 2006) and for the selective dissemination of information in digital libraries (Peis *et al*., 2008).

In this study we deal with the use of syndication in the library field, analyzing the possibilities of transmission and retrieval of textual bibliographic catalogs using content syndication.

MATERIALS AND METHODS

Methodology and results: We had first to develop test collections of different sizes from the bibliographic catalog of the LC (2012) through very general topic queries. Their names and dimensions can be seen in the Table 1:

Once the collections were in CSV format, a program (developed in PHP) converted them into raw XML format. The collections were split in groups of one thousand records each to improve the later input speed of data in My SQL. We call this process data conversion. The Table 2 summarizes the conversion times obtained:

Then the several collections were syndicated in various formats (ATOM, RSS 1.0 and RSS 2.0), to compare the performance of different syndication formats, adding another one (MARC-XML with two variants: short and extended) not currently considered as a syndication format, though widely used in the library world.

A structure for each format was created. To do it we had chosen all the labels that, being used in practice, were more useful to describe textual (not multimedia) bibliographic records, avoiding possible loss of information. In the case of the MARC format it was considered bibliographic corpus consisting mainly of monographs and the use of the Dewey decimal classification, as the collection source was from the Library of Congress. The structure chosen for the ATOM format (The ATOM Syndication Format, 2005) is as Fig. 1.

Tags chosen to set the channel in RSS 1.0 syndication format (RDF Site Summary 1.0, 2008), along with the specifications of the modules included (RDF Site Summary 1.0 modules: Content, 2002; RDF Site Summary 1.0 modules: Dublin Core, 2000; RDF Site Summary 1.0 modules: PRISM, 2002; RDF Site Summary 1.0 modules: SKOS, 2009), are as Fig. 2.

Corresponding Author: Manuel Blazquez Ochando, Faculty of Library and Information Sciences, University Complutense, Santísima Trinidad 37, 28010 Madrid Spain

ATOM

```
<entry>
<id>Identifier number</id>
<title>Title area</title>
<author><name>Statement of responsibility area</name></author>
<updated>Update date of the bibliographic record</updated>
<content>Full bibliographic record</content>
<link rel='alternate' href= 'Permanent URL of the bibliographic record'/>
<summary>Abstract of content</summary>
<category term='Topic of the document using keywords or classification systems'/>
<contributor><name>Other statement of responsibility</name></contributor>
<published>Publication area</published>
<source>Channel URL of the original record</source>
<rights> Rights about the bibliographic record </rights>
</entry>
```

Fig. 1: Bibliographic record structure in ATOM format

RSS 1.0

```
<item>
<dc:title> Title area </dc:title>
<dc:creator> Statement of responsibility area </dc:creator>
<dc:contributor> Other statement of responsibility </dc:contributor>
<dc:publisher> Publication area </dc:publisher>
<dc:date>Publication date</dc:date>
<dc:type>Material or type of resource area (text, image, sound) and nature of the document
(according to control vocabulary, for example: monographs, periodicals, serials…)</dc:type>
<dc:format> MIME type format document, size or duration </dc:format>
<dc:identifier> Unique identifier of the document or permanent URL in bibliographic record
</dc:identifier>
<dc:subject> Topic of the document using keywords or classification systems </dc:subject>
<dc:source> Channel URL of the original record </dc:source>
<dc:language>Language of document</dc:language>
<dc:relation>Related contents (for example: collection of documents, other sources and related
resources)</dc:relation>
<dc:coverage>Temporal or spatial coverage of the content</dc:coverage>
<dc:rights> Rights about the bibliographic record </dc:rights>
<prism:publicationName>Title area of the periodical</prism:publicationName>
<prism:edition>Edition area of the periodical</prism:edition>
<prism:publisher>Publisher of the periodical</prism:publisher>
<prism:publicationDate>Publication date of the periodical</prism:publicationDate>
<prism:issn>ISSN</prism:issn>
<skos:note>Content note</skos:note>
<content:encoded> Full bibliographic record </content:encoded>
</item>
```

Fig. 2: Bibliographic record structure in RSS 1.0 format

RSS 2.0

```
<item>
<title> Title area </title>
<author> Statement of responsibility area </author>
<pubDate>Publication date</pubDate>
<source> Channel URL of the original record </source>
<guid> Unique identifier of the document </guid>
<link>Permanent URL of the bibliographic record</link>
<description> Abstract of content </description>
<enclosure>URL, size and MIME type of the original document attached </enclosure>
<comments>URL of the comments webpage and ranking or review of the original
document</comments>
<category> Topic of the document using keywords or classification systems </category>
</item>
```

Fig. 3: Bibliographic record structure in RSS 2.0 format

Table 1: Characteristics of the collections crested

	Size (MB)	Number of records
1000_records	0.77	1001
5000_ records	2.68	5002
10000_ records	2.05	10004
25000_ records	13.33	25008
50000_ records	28.34	50036
100000_ records	54.95	100054
250000_ records	144.00	250146
500000_ records	280.49	500309
1000000_ records	561.39	1000039

From the specifications of RSS 2.0 format (RSS 2.0 specification, 2003) the structure has been configured as Fig. 3.

To set records in MARC-XML format, the specifications of the Library of Congress and the MARC Standards Office (MARC-XML schema, 2009) were used. The structures of the short and extended versions are shown in Fig. 4 and 5 respectively:

MARC 1 (short)

```
<marc:record>

<controlfield tag='001'>Control Number </controlfield>
<controlfield tag='003'>Control Number Identifier </controlfield>

<datafield tag='017' ind1=" ind2="">
<subfield code='a'>Copyright or Legal Deposit Number</subfield>
</datafield>

<datafield tag='020' ind1=" ind2="">
<subfield code='a'>ISBN International Standard Book Number</subfield>
</datafield>

<datafield tag='022' ind1='0' ind2="">
<subfield code='a'>ISSN International Standard Serial Number</subfield>
</datafield>

<datafield tag='035' ind1=" ind2="">
<subfield code='a'>System Control Number</subfield>
</datafield>

<datafield tag='041' ind1='0' ind2="">
<subfield code='a'>Language Code</subfield>
</datafield>

<datafield tag='043' ind1=" ind2="">
<subfield code='c'>Geographic Area Code</subfield>
</datafield>

<datafield tag='082' ind1=" ind2="">
<subfield code='a'>Dewey Decimal Classification Number</subfield>
</datafield>

<datafield tag='100' ind1='1' ind2="">
<subfield code='a'>Main Entry - Personal Name</subfield>
</datafield>

<datafield tag='245' ind1='1' ind2="">
<subfield code='a'>Title Statement </subfield>
</datafield>

<datafield tag='250' ind1=" ind2="">
<subfield code='a'>Edition statement</subfield>
<subfield code='b'>Remainder of edition statement</subfield>
</datafield>

<datafield tag='260' ind1=" ind2="">
<subfield code='a'>Place of publication, distribution, etc.</subfield>
<subfield code='b'>Name of publisher, distributor, etc. </subfield>
<subfield code='c'>Date of publication, distribution, etc. </subfield>
</datafield>

<datafield tag='300' ind1=" ind2="">
<subfield code='a'>Physical Description , extent, size</subfield>
</datafield>

<datafield tag='310' ind1=" ind2="">
<subfield code='a'>Current Publication Frequency </subfield>
</datafield>

<datafield tag='490' ind1='0' ind2="">
<subfield code='a'>Series statement </subfield>
<subfield code='v'>Volume/sequential designation </subfield>
</datafield>

<datafield tag='500' ind1=" ind2="">
<subfield code='a'>General Note </subfield>
</datafield>

<datafield tag='654' ind1='0' ind2="">
<subfield code='a'>Subject Added Entry - Faceted Topical Terms </subfield>
</datafield>

</record>
```

Fig. 4: Bibliographic record structure in MARC-XML format (short)

From the collections shown in Table 1, a program in PHP creates the channels in different formats with previous structures. Creation times obtained are as follows Table 3.

Once all collections (from 1000 up to 1 million records) were syndicated in different formats, the diffusion process of data in a channel from the server to the client computer is simulated by an import program (also developed in PHP) whose main goals are:

- Identify the format of the channel through its extension
- Create a data table in MySQL with a structure adapted to the syndication format
- Sequential reading of each bibliographic record. The time spent in this process is the transfer time between the server and the client computer. This transfer time in practice depends on many factors, including the bandwidth of the network, the processing speed of the client computer or system memory. Therefore this time has not been considered in this study

MARC 2 (extended)

```
<marc:record>

<marc:controlfield tag='001'>Control Number </marc:controlfield>
<marc:controlfield tag='003'>Control Number Identifier </marc:controlfield>

<marc:datafield tag='017' ind1=" ind2=">
<marc:subfield code='a'>Copyright or Legal Deposit Number</marc:subfield>
</marc:datafield>

<marc:datafield tag='020' ind1=" ind2=">
<marc:subfield code='a'>ISBN International Standard Book Number</marc:subfield>
</marc:datafield>

<marc:datafield tag='022' ind1='0' ind2=">
<marc:subfield code='a'>ISSN International Standard Serial Number</marc:subfield>
</marc:datafield>

<marc:datafield tag='035' ind1=" ind2=">
<marc:subfield code='a'>System Control Number</marc:subfield>
</marc:datafield>

<marc:datafield tag='041' ind1='0' ind2=">
<marc:subfield code='a'>Language Code</marc:subfield>
</marc:datafield>

<marc:datafield tag='043' ind1=" ind2=">
<marc:subfield code='c'>Geographic Area Code</marc:subfield>
</marc:datafield>

<marc:datafield tag='082' ind1=" ind2=">
<marc:subfield code='a'>Dewey Decimal Classification Number</marc:subfield>
</marc:datafield>

<marc:datafield tag='100' ind1='1' ind2=">
<marc:subfield code='a'>Main Entry - Personal Name</marc:subfield>
</marc:datafield>

<marc:datafield tag='245' ind1='1' ind2=">
<marc:subfield code='a'>Title Statement </marc:subfield>
</marc:datafield>

<marc:datafield tag='250' ind1=" ind2=">
<marc:subfield code='a'>Edition statement</marc:subfield>
<marc:subfield code='b'>Remainder of edition statement</marc:subfield>
</marc:datafield>

<marc:datafield tag='260' ind1=" ind2=">
<marc:subfield code='a'>Place of publication, distribution, etc.</marc:subfield>
<marc:subfield code='b'>Name of publisher, distributor, etc. </marc:subfield>
<marc:subfield code='c'>Date of publication, distribution, etc. </marc:subfield>
</marc:datafield>

<marc:datafield tag='300' ind1=" ind2=">
<marc:subfield code='a'>Physical Description , extent, size</marc:subfield>
</marc:datafield>

<marc:datafield tag='310' ind1=" ind2=">
<marc:subfield code='a'>Current Publication Frequency </marc:subfield>
</marc:datafield>

<marc:datafield tag='490' ind1='0' ind2=">
<marc:subfield code='a'>Series statement </marc:subfield>
<marc:subfield code='v'>Volume/sequential designation </marc:subfield>
</marc:datafield>

<marc:datafield tag='500' ind1=" ind2=">
<marc:subfield code='a'>General Note </marc:subfield>
</marc:datafield>

<marc:datafield tag='654' ind1='0' ind2=">
<marc:subfield code='a'>Subject Added Entry - Faceted Topical Terms </marc:subfield>
</marc:datafield>

</marc:record>
```

Fig. 5: Bibliographic record structure in MARC-XML format (Extended)

- Insert each bibliographic record (in groups of one thousand, having been found that this level of clustering minimizes the insertion time) in the data table. This process has been called data insertion. In this study, since transfer time has not been considered, insertion time represents the total import time of syndicated sources

The results obtained in this process, considering the different collections and different formats are summarized in the following Table 4:

RESULTS AND DISCUSSION

We see that each format has a different capacity for the syndication of bibliographic catalogs. This is due to a different internal structure, which allows or denies the input of certain types of information from the bibliographic record.

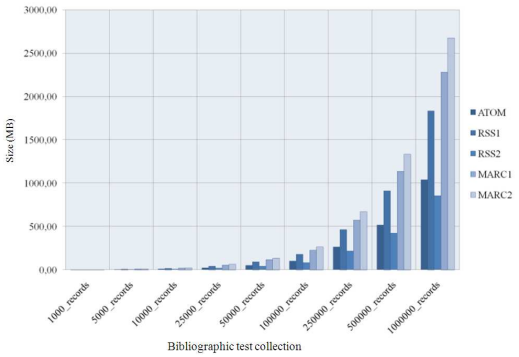

Fig. 6: Syndicated collections size

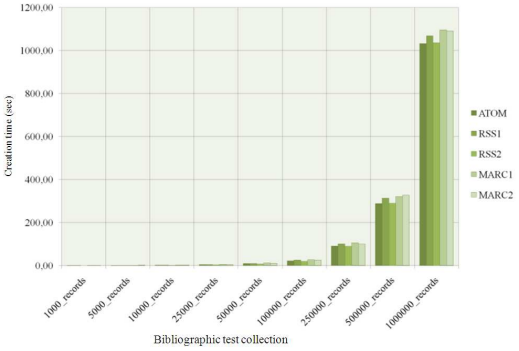

Fig. 7: Creation time of Syndicated channels

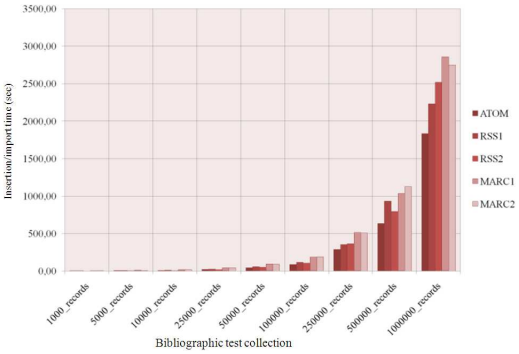

Fig. 8: Insertion/import time of Syndicated channels

From a librarian point of view, a bibliographic record consists of the following essential information areas: title and statement of responsability, edition, material or type of resource, publication, physical description, series and notes (IFLAI, 2007). We have added the possibility to include the full bibliographic record, because it would help to retrieve any information or specific secondary access point that would not have been considered in the original format. The following Table 5 summarizes the information areas that each format can include:

Table 2: Conversion time from CSV to XML raw

Collection	Conversion time (sec)
1000_records	1.21
5000_ records	4.93
10000_ records	9.45
25000_ records	24.54
50000_ records	50.11
100000_ records	99.13
250000_ records	251.91
500000_ records	504.23
1000000_ records	992.36

Table 3: Creation time of channels

Format	Collection	Generation time (sec)
ATOM	1000_records	0.19
	5000_ records	0.74
	10000_ records	1.32
	25000_ records	3.64
	50000_ records	8.52
	100000_ records	20.90
	250000_ records	90.89
	500000_ records	287.61
	1000000_ records	1032.25
RSS 1.0	1000_records	0.17
	5000_ records	0.65
	10000_ records	1.65
	25000_ records	4.34
	50000_ records	8.80
	100000_ records	24.59
	250000_ records	100.34
	500000_ records	312.98
	1000000_ records	1068.93
RSS 2.0	1000_records	0.11
	5000_ records	0.94
	10000_ records	1.21
	25000_ records	3.37
	50000_ records	8.13
	100000_ records	20.28
	250000_ records	89.51
	500000_ records	290.36
	1000000_ records	1036.30
MARC-XML 1	1000_records	0.24
Abbreviated	5000_ records	0.81
	10000_ records	1.75
	25000_ records	4.82
	50000_ records	11.78
	100000_ records	26.68
	250000_ records	89.51
	500000_ records	290.36
	1000000_ records	1036.30
MARC-XML 2	1000_records	0.20
Abbreviated	5000_ records	0.84
	10000_ records	1.70
	250000_ records	99.03
	500000_ records	326.96
	10000000_ records	1091.31

This Table 5 shows that the most suitable format for bibliographic records syndication is the family of MARC formats, as they can include all the essential areas of bibliographic description with the level of completeness desired. This possibility of complete full bibliographic description justify why it is not allowed to add the full bibliographic record, as it would duplicate the information.

Nevertheless, it would allow displaying the full record and the access to any data in a single field if it was necessary.

Between syndication formats, RSS1 is the best suited to a textual bibliographic record, mainly because of the possibility of including Dublin Core and PRISM modules. However several shortcomings have been detected, such as the impossibility of including areas of publication, physical description and series. These problems can be overcome at least partially due to the possibility of including a full bibliographic record field on which to add such data.

RSS2 and ATOM have a similar level of adaptability. Both have a low capacity to represent a textual bibliographic record, basically limited to the area of title and statement of responsibility. RSS2 presents the added disadvantage of not including the full bibliographic record. This problem can be solved, but in an unorthodox way, by adding modules initially designed for the RSS1 format (Dublin Core and PRISM) after introducing the respective namespaces. This solution would originate a hybrid format far away from the original, becoming a substitute of RSS1.

Another relevant aspect has to do with the time of creation of syndicated channels, analyzing its relationship with the size of the files in the different formats. To observe this relationship, first we show in the Fig. 6 the sizes of the syndicated files.

As shown in the graph, a difference in the structure of the formats affects the size of the channels. Logically, the more complex and extensive is the structure, the bigger is the size of the channel. Formats range from the simplest, RSS2, to the most complex, MARC-XML extended.

However, the differences in the size of the channels don't imply large differences in the time of creation of such channels, as shown in the Fig. 7.

In fact, although the one million bibliographic records syndicated in MARC-XML extended becomes a 2673 MB channel and the same collection in RSS2 becomes a 854 MB channel (three times less), the creation times are 1091 and 1036 sec respectively, which is around 1 min of difference. Even more, as the size of the initial collection reduces, the differences in creation times also reduce, as can be seen in Table 3. In summary, the format (despite the structural differences between them) does not have a major effect on the time of creation of syndicated channels.

In relation to the import times (considering only the insertion time into the database), the corresponding Table 4 can be summarized graphically as follows:

The Fig. 8 shows that the difference in insertion times between the different formats is high when the initial collections are big, but this insertion time decreases as the collection becomes littler.

Table 4: Insertion import time

Format	Collection	Data entry time (sec)
ATOM	1000_records	0, 75
	5000_ records	3, 10
	10000_ records	5, 27
	25000_ records	17, 24
	50000_ records	42, 35
	100000_ records	86, 27
	250000_ records	284, 09
	500000_ records	630, 23
	1000000_ records	1832, 74
RSS 1.0	1000_records	1, 05
	5000_ records	3, 45
	10000_ records	6, 55
	25000_ records	21, 27
	50000_ records	54, 73
	100000_ records	113, 13
	250000_ records	351, 60
	500000_ records	930, 99
	1000000_ records	2229, 34
RSS 2.0	1000_records	0, 45
	5000_ records	3, 45
	10000_ records	5, 83
	25000_ records	20, 21
	50000_ records	50, 41
	100000_ records	105, 17
	250000_ records	363, 52
	500000_ record	794, 03
	1000000_ records	2519, 13
MARC-XML 1 Abbreviated	1000_records	1, 68
	5000_ records	8, 36
	10000_ records	16, 85
	25000_ records	42, 63
	50000_ records	92, 88
	100000_ records	184, 64
	250000_ records	510, 92
	500000_ records	1034, 99
	1000000_ records	2857, 61
MARC-XML 2 Abbreviated	1000_records	4, 05
	5000_ records	8, 45
	10000_ records	17, 21
	25000_ records	43, 11
	50000_ records	92, 56
	100000_ records	188, 49
	250000_ records	508, 32
	500000_ records	1125, 76
	1000000_ records	2749, 38

Table 5: Information area of bibliographic records and syndication formats

	ATOM	RSS 1	RSS 2	1-MARC 1	2-MARC 2
Title and statement of responsibility area	X	X	X	X	X
Edition area				X	X
Material type of resource area		X		X	X
Publication area	X	X		X	X
Physical description area				X	X
Series area				X	X
Notes area		X		X	X
Resource identifier	X	X	X	X	X
Full bibliographic record	X	X			

For the collections of a million records, the maximum difference obtained (between MARC extended format and ATOM format) is around 15 min. For the collections of half a million records, the maximum difference is approximately 8 min between the same formats. For the collections of 250,000 records, the maximum difference is around 4 min. This maximum difference is 1.7 min with the collections of 100,000 records, less than a minute (50.53 sec) in the case of 50,000 records and finally it is 3.3 sec in the case of 1000 entries.

In summary, although the time of creation of the channels does not almost depend on the size of collections, the time of insertion/import strongly depends on the size of collections. The data obtained let us conclude that syndicated data insertion/import is less than a minute, regardless of the format chosen, only when the collection does not exceed 25,000 records. For collections equal or greater than 250,000 records, the insertion/import times are considerably large.

CONCLUSION

The analysis suggests that content syndication is a useful technique for the transmission and retrieval of textual bibliographic data, being an alternative to the use of Z39-50 protocol, more complex and difficult to use. The smaller is the syndicated collection of records, the more useful is this technique. When the collection is smaller than 25,000 records, the insertion/import time is less than a minute, regardless of the format chosen. Accordingly, this technique is well suited for updating library catalogs or for the maintenance or management of large databases that are fed from multiple sources.

MARC-XML has been shown to be the most complete format, because it has specific tags for any bibliographic data. RSS1 has also shown useful and versatile for the representation of bibliographic records due to the inclusion of various specialized modules of description, such as Dublin Core and PRISM. Although it lacks the areas of publication, physical description and series, it has an area of content that lets you insert the full bibliographic record, overcoming these deficiencies. In any case, this tag would help to display and retrieve any kind of bibliographic information, regardless of the format used. According to these criteria, RSS2 and ATOM are the less suitable syndication formats for the transfer of bibliographic data.

The analysis suggests that there are no appreciable differences in the time of creating the channel, whatever the chosen format and the complexity of its structure. However, we have found that the greater is the complexity of the format, the greater is the size of the channel, implying an increase in the insertion/import time. Thus, although it is technically feasible to

syndicate collections of any size, only with collections smaller than 25,000 records the insertion/import time is less than one min.

Further research directions: Further research could analyze the syndication of bibliographic or library collections in real network environments, primarily to determine what factors influence primarily on their performance. In this environment it could be possible to compare their performance with Z39.50 protocol.

Another direction for future research is the development of techniques for retrieving information over syndicated library collections through XQuery or XPath filtering techniques and their comparison with the usual MYSQL techniques.

REFERENCES

Abadal, E., A. Estivill, J. Franganillo, J. Gascón and R.J.M. Gairín, 2006. Sindicación de contenidos en un portal de revistas: Temaria. El Profesional de la Información, 15: 214-221.

ANU, 2010. The Australian National University.

Brickley, D. and R.V. Guha, 2004. RDF vocabulary description language 1.0: RDF schema.

IFLAI, 2007. International Standard Bibliographic Description (ISBD). 1st Edn., Walter de Gruyter, Mϋnchen, ISBN-10: 3598242808, pp: 322.

LC, 2012. Library of Congress online catalog. Library of Congress.

MARC-XML schema, 2009.

New York Times, 2009. The New York Times news service/syndicate.

Peis, E., E. Herrera-Viedma and J.M. Morales-del-Castillo, 2008. Modelo de servicio semántico de difusión selectiva de información (DSI) para bibliotecas digitales. El Profesional de la Información, 17: 519-525.

RDF Site Summary (RSS) 1.0. 2008.

RDF Site Summary 1.0 modules: Content 2002.

RDF Site Summary 1.0 modules: Dublin Core, 2000.

RDF Site Summary 1.0 modules: PRISM, 2002.

RDF Site Summary 1.0 modules: SKOS, 2009.

RSS 2.0 specification, 2003.

The ATOM Syndication Format, 2005.

Winer, D., 2001. Scripting news. Dave Winer.