

TICTOCRON: AN AUTOMATIC SOLUTION FOR PROPAGATING QUALITY METADATA TO SCHOLARLY TOC RSS FEED METADATA

Santiago Chumbe¹ and Roddy MacLeod²

^{1,2}Heriot Watt University, UK

¹S.Chumbe@hw.ac.uk, ²R.A.MacLeod@hw.ac.uk

Abstract

Institutions and researchers stand to benefit from the facilitation of more widespread syndication of, and easier access to, Table of Content (TOC) RSS (Really Simple Syndication [1]) feeds produced for scholarly journals. However, many journal TOC RSS feeds are at present being produced with erroneous, poor or incomplete metadata. This can hamper the usefulness of scholarly current awareness services, and also cause problems for individual subscribers to those feeds. This is exactly what the ticTOCron software toolkit aims to overcome. The ticTOCron toolkit automatically enhances poor, heterogeneous and incomplete metadata found in TOC RSS feeds by making use of a pre-defined "Best Practice" metadata scheme suitable for scholarly journals. In this work we depict the main issues and "bad practices" found in TOC RSS metadata obtained from more than 435 scholarly publishers. Then, we describe software solutions implemented via ticTOCron. Some references are made to the algorithms for generating semantic relations within, between and from the harvested TOCs and to the mechanisms for propagating "metadata associations" from a previously crawled metadata-rich reference set. However, an effort is made to avoid technical jargon and to replace complex technical descriptions with samples and simple comparisons. The original metadata is converted to a canonical format using the "Best Practices metadata set" for scholarly papers proposed by the ticTOCs Project [2]. We also present the results produced by ticTOCron when it was used for enhancing and normalizing TOC RSS feeds collected from more than 12,000 journals. Finally we propose a sustainable and scalable computational model whereby the automatic solution is complemented and fine-tuned by a cost-effective human cross-validation process.

Keywords:

Metadata quality enhancement, Journal TOC RSS feeds, Current Awareness, CRON job

1. Introduction

Keeping up-to-date with the latest publications is at the heart of much research activity. Publishers, librarians, authors and readers have an interest in ensuring better exploitation and early access to the latest output research. That is why considerable effort is being expended by a growing number of publishers on providing RSS feeds for their journal TOCs.

Using RSS readers, users can view details of the latest articles as soon as they have been published, without having to visit individual publishers' websites where the TOC has been taken from. Clearly, all stand to benefit from the facilitation of more widespread syndication and easier access to TOC RSS feeds. RSS readers and current awareness services that merely propagate what publishers have put in their RSS feeds do not solve the key problems that researchers face when using TOC RSS feeds to discover up-to-date knowledge from an avalanche of ever increasing research output.

Achieving an automatic mechanism to serve the needs for currency of this growing number of users has been always challenging. Online current awareness services are important to virtually all librarians and researchers as has been noticed by Farooq [Farooq, U. et al 2008]. As part of this process, users have a growing interest in the RSS feeds that publishers are providing for an increasing number of scholarly journal's TOCs. However, we have found that the vast majority of publishers are including erroneous, insufficient or poor quality metadata in their TOC RSS feeds. The reasons why publishers produce poor quality TOC RSS feeds can vary from lack of

knowledge about the technology or its benefits, to being reluctant to implement expensive metadata quality assurance systems.

- More than **6 million researchers** worldwide
- Around **1.5 million articles** are written annually
- **About 24,000 scholarly journals**
- **2,000 publishers** publishing these scholarly journals
- The number of published journals and researchers increases by about 3% per annum.
- More than **16,000** journals provide TOC RSS feeds

(Various sources, including the International Association of Scientific, Technical and Medical Publishers)

To help to solve or mitigate this problem, ticTOCs on one hand has prepared and proposed recommendations for publishers to understand the importance of providing quality metadata in their feeds. On the other hand ticTOCs has developed a semi-automatic mechanism for augmenting the quality of metadata harvested from TOC RSS feeds. The combined effect of these two works has enabled the prototyping of a web-based environment that allows researchers to discover, subscribe to, personalise, export and reuse TOC RSS feeds. It also supports the creation of APIs (Application Programming Interfaces) to facilitate the re-use of aggregated journal TOC content on a subject or community basis by gateways, subject-based resource discovery services, library services, portals and other services.

Section two of this paper will discuss the deficiencies identified in the TOC RSS feeds. Most of those problems, such as parsing erroneous XML files can be solved by using well-known software tools such as the Universal Feed Parser [Lerner, 2004]. However, the lack or insufficiency of desirable information in the TOC RSS feeds is a problem that so far has not been solved systematically. RSS feeds are usually just consumed but not augmented by RSS readers. In this paper we describe a solution that combines metadata crawled from the web with normalization of metadata extracted from TOC RSS feeds. We use content found on the web to complete the “missing” information in the TOC RSS feeds. In fact what we have done is to generate “new” metadata from related pre-existing metadata sources. In particular our development has been inspired by the work done by Lagoze [Lagoze et al, 2006] who has reported using web crawler-based metadata to augment the quality of poor OAI (Open Archives Initiative) metadata with significant positive results.

In the next sections we describe the algorithm and the composition of the ticTOCron toolkit used by the Project to automatically collect (harvest) TOC RSS feeds and augment their quality. We conclude by discussing the impact of ticTOCron on the quality of TOC RSS feeds, its possible applications for supporting or enhancing the work of specific communities of users, and finally we draft some conclusions.

2. Problems with Journal TOC RSS Feeds

There are problems associated with the technology used to produce the feeds. Other major problems are associated with the quality and the content of the feeds. Although more than 80% of TOC RSS feeds provide the TOC for the latest issue of the journal, the variations in the quality and type of content are vast, and range from simple alerts that a new issue is available, to whole back-catalogues of issues. We have classified those problems in three main categories.

2.1. Technology Barriers

RSS is another example of theoretical standards undermined by chaotic reality of the web. To make things worse, RSS has many competing standards (in practice, Atom is just another syndication standard [Hammersley, 2005]). There are two main branches of RSS: RSS 1.0 and RSS 2.0 [Hammersley, 2005]. RSS 2.0 format is simpler than RSS 1.0, which can be extended by the use of modules [Wittenbrink, 2005] to provide rich metadata. We found out that only two

publishers were exclusively using Atom to provide TOC feeds. Nevertheless, handling different syndication formats is no longer an issue for service providers.

Feed Standard	Journals	%
RSS 2	6,598	53.7
RSS 1 (RDF)	5,626	45.6
Atom	6	0.0
XML files	91	0.7
Total:	12,321	100.0

Table 1. Number of journals using different syndication formats (as Jan 2008.)

RDF Module	Journals
- (<i>no modules</i>)	4,506
dc (Dublin Core)	3,433
dc prism	3,966
dc prism content	401

Table 2. Number of journals using RDF modules recommended by ticTOCs (as Jan 2008.)

Invalid and non well-formed TOC RSS feeds are also a common issue. Thus, it is quite common to find feeds that use HTML mark-up within the title or description tags. Some publishers are not aware of modules such as *content:encoded*, which is a suitable way to display HTML including hyperlinks to images. In general, publishers are failing to notice that the metadata included in their TOC RSS feeds have the potential of being more than simple headlines. Without a focus on re-use, TOC RSS metadata is less useful. With that aim, we have encouraged publishers to use a set of available RDF modules [5] to enrich their feeds. Currently there are 3 official modules and 19 proposed modules [Manola et al, 2004]. Any syndication format can be enriched with RDF modules, however use of RSS 1.0 would be advisable as it follows the RDF specifications. The *PRISM*, *Syndication*, *Content* and *Dublin Core* modules [5] have been identified as suitable modules for TOC RSS feeds [Rogers, 2008].

However, as RSS 1.0 is extensible by design, modules can be written by anyone, which can unnecessarily produce heterogeneous metadata encodings. For example, the following “unsuitable” modules were found in some TOC RSS feeds collected from publishers: CC, FEEDBURNER, FOAF, and HTML. Yet, it was encouraging to notice that by Feb 2009, more than 2,000 TOC RSS feeds have started to include elements from the *Syndication* module in their feed’s headers. This is a positive development. The *Syndication* module has the potential to enable toolkits such as ticTOCron with the capacity of performing selective harvesting, which would save computational time and make it possible to update only the new TOC RSS feeds and as soon as they have been published.

2.2. Metadata tagging inconsistencies

There is a widespread difference in what publishers are putting in their feeds. For further information there are the deliverables of the ticTOCs Project [2].

The conclusion expressing that “*you cannot trust the metadata provided by any OAI data provider*” [Chumbe, 2006] also applies to TOC RSS feeds providers. The values that publishers are including in metadata elements could be almost anything, regardless of the original purpose of the element specified by metadata standards. For example Elsevier put the authors of the article in the <description> element; Springer-Verlag uses <description> to store the title, abstract, authors, vol, issue, year, DOI, etc.; Biomed Central puts all the authors in a single <dc:creator> element; and IoP uses a <dc:creator> element for each author.

- multiple authors – IOP:

```
<dc:creator>Luciano Telesca</dc:creator>
<dc:creator>Antonio Lanorte</dc:creator>
<dc:creator>Rosa Lasaponara</dc:creator>
```

- multiple authors - Biomed Central:

```
<dc:creator>Alexandra Devine, Michelle Kermode, and Helen Herman</dc:creator>
```

- multiple authors – Elsevier:

```
<description>Dostal, M. , Roberts, J.B. , Holmes, R.</description>
```

- multiple authors – Springer-Verlag:

<description><html tags>article title, abstract, authors, Vol, Issu, Year, ISSN, etc..</html tags></description>

In general, any value provided in the metadata is a potential source of error. Odd values have been identified in the feeds, such as “<date>2086</date>”. CrossRef [3], a main partner of ticTOCs is expected to play a leading role in encouraging every publisher “*to play ball according to the rules.*”

2.3. Feeds Content inconsistencies

Describing all the inconsistencies found in the content of TOC RSS feeds would take too long. The inconsistencies are varied. Some journals, such as the “Drugs in R & D” Journal, puts all the metadata in the RSS feed header. Other journals use different types of formats to register dates. The following is a summary of common inconsistencies detected in the content included in TOC RSS feeds.

- Various TOCs in the same feed. For ticTOCs, being a current awareness service, the back-catalogue of issues previously published in a journal is irrelevant, and this information might also confuse individual RSS feed subscribers.
- Some TOC RSS feeds contain other, different types of content. Not only TOCs, but also News, Job Announcements, Editorials, etc.
- Very few journals use the proper metadata elements to register information on Vol, Issue or Year. A significant number of feeds do not include this information at all.
- Because in RSS feeds there isn’t such a concept as “mandatory” it is quite common to find feeds without enough basic information such as the title’s item and the link to its webpage.

UTF-8 encoding errors are also common. RSS Feeds generated by software can be an obstacle for automatic harvesting. Thus <https://pi.library.yorku.ca/ojs/index.php/soi/feed/rss> redirects harvesters to an authentication webpage. The above examples are just some cases illustrating poor metadata being exposed by TOC RSS feeds.

3. The ticTOCron toolkit

The aim of the toolkit is to automatically harvest TOC RSS feeds and to use a suite of software tools to handle the problems affecting TOC RSS feeds as identified in the previous section. The toolkit works as an “invisible infrastructure” for the ticTOCs public service prototype. ticTOCron itself is a set of software tools managed by a computer system utility (CRON job) found in most UNIX systems [Aeleen, 2002].

The composition of the ticTOCron software is based on an adaptation of the classification proposed by Berson [Berson, A. et al 2007] to handle the gathering of metadata that we need to consume but for whose quality we don’t have any influence. Berson classifies in five categories the tools that can partially or completely automate the tasks associated with extracting, cleaning, standardizing and enriching metadata obtained from external sources:

- Auditing Tools
- Data Cleansing Tools
- Data parsing and standardization Tools
- Data extraction, transformation and loading tools
- Hybrid Packages

Each of the ticTOCron tools is a modular software component that has been developed using a loosing coupling design technique. Each of the tools has a specific goal as outlined by Bearson. Therefore, the elimination or failure of any of these tools does not bring the entire CRON system down. The sequence in which they are executed is computationally irrelevant, but it has an impact on the consistency of the results produced by the CRON process.

The tools produce logs and exception errors that are used by the system administrator to solve errors and more importantly to increase the “*know-how*” of the tools.

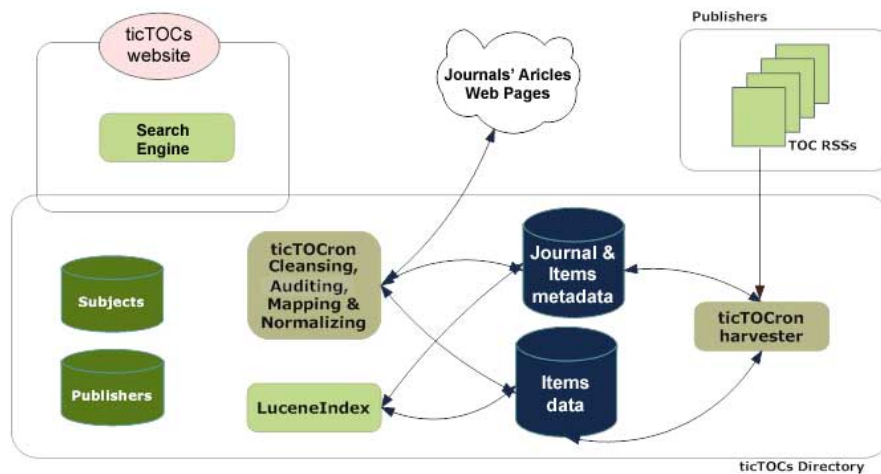


Figure 1: ticTOCron toolkit system context

3.1 Harvesting tool

There is a large number of software tools for parsing RSS feeds but most of them will run into problems associated with invalid and/or non well-formed XML files. Those RSS readers handle faulty feeds with different levels of sophistication. Some of them will attempt to correct the faulty feeds, but others will simply return an error in the user interface. However a basic requirement for our harvester was that it should be able to parse even non well-formed XML files and it should not be stopped by unpredictable XML content. Compliance with RSS standards is not taken into account at this stage. The tool will accept any XML file and if the file is not a valid RSS feed, will intend to identify any TOC-like structure in the file and try to extract metadata. If the process fails, the harvester will log the event and continue parsing the next RSS feed.

```

Select journals that haven't been updated in the last week, starting with the
ones that are more frequently updated or their frequency is unknown
For each selected journal
{
  Check frequency of publication if available
  If it needs to be updated or frequency is not available
  {
    Harvest journal TOC RSS feed
    If there is a problem, log error and continue with next journal
    Generate and update metadata for the journal feed (e.g. RSS version,
used modules, frequency of publication, etc.)
    Parse feed to compare its contents with what is stored in the
database
    If contents have not changed continue with next journal
    Remove items from previous feed
    For each new item
    {
      Save item data
      Generate and store additional common metadata, such as
quality level (0=need normalization, 1=need cleansing,
2=need auditing,
3=need cleansing & auditing. Default is 0)
    }
  }
}

```

Figure 2. ticTOCron Harvester Tool Algorithm

As most of the journal feeds do not provide the frequency of publication, the harvesting process takes a long time to complete. The method is still far from perfect because in order to succeed it requires having a long enough harvesting history for each journal. The computing time is inversely proportional to the number of feeds whose frequency of publication is known.

Despite its importance, the automatic collection or harvesting of RSS feeds is the easiest task done by the CRON job. The challenging part starts after the metadata has been collected and deposited in the database.

3.2. Data Cleansing Tool

This utilises a set of rules that have been setup in the software to remove, combine or move harvested values to improve the quality of the metadata and potentially to add new accurate content. It deals with the TOC RSS items whose quality level is one (please see Fig. 2.)

Basically, the Cleansing tool analysis the value of each metadata element with two aims; firstly to remove irrelevant items to leave only the ones of the current issue; and secondly, to move or/and create metadata from the harvested information. Thus, the tool will remove from the TOC all the items that have unsuitable information such as news, job announcements as well as old issues. Then it will try to identify relevant missing information from each metadata element for each TOC item.

For example if the Cleansing tool detects the presence of a valid DOI (Document Object Identifier) in a metadata element, it will create a new metadata element for the DOI. In most of the cases, the tool can identify values commonly used for citing a scholarly journal article. However, the tool will be unable to detect relevant information put by the publisher in the wrong metadata element if none of its rules is matched by contents of this element. For example we have noticed that the feeds using the PRISM module are the ones that require little cleansing work. The most challenging cases are produced by feeds that store different types of useful information in the wrong element. An extreme case is illustrated by some TOC RSS feeds produced by Springer-Verlag. The description field of those feeds contains the article abstract plus all the information that is required to create the full article citation (Journal Title, ISSNs, article title, abstract, authors, Vol, Issue, publication year, URLs, DOI, etc.) and all of them enclosed by HTML tags.

The Cleansing tool analyses all the available information available in the TOC items looking for additional information to augment the metadata quality and richness. It is also in charge of handling the correct metadata encoding (UTF-8 conversion) as well as of removing html mark-up and spurious tags from the TOC items.

3.3. Auditing Tool

This tool compares the metadata of each TOC RSS item against the item's metadata found in the publisher website. The objective is to enhance the accuracy of the RSS metadata and its correctness by contrasting the metadata values of each item of the TOC RSS feed against information crawled from the item's web page. The success of the Auditing tool depends on its ability to identify, in the item's web page, metadata that is better than the TOC RSS metadata. This factor in turn depends on the tool's algorithm and the structure of the web page that has been crawled and is being analysed.

There is a wide range of software tools for creating metadata automatically from crawled web pages [Albassuny, 2008; Liu, 2007]. Our tool is based on heuristic models that use the formatting and composite structure of web pages rendering scholarly articles metadata to break the ambiguousness of HTML documents [Tonkin, 2008]. Additionally, taking into account that extraction algorithms could contribute to useful automatic metadata generation [Greenberg, 2003], the Auditing tool uses a standard HTTP GET request to download the item's webpage and utilises its own TOC RSS metadata to create associations between the crawled metadata values and the metadata values obtained from the RSS feeds; then the problem of extracting metadata from the web can be reformulated as one of inter-extrapolation of values between a previously crawled metadata-rich reference set and a related metadata-poor source (RSS.) The

knowledge gained in each extraction is stored and used as a “reference data model” for the next time that the “same” web page is parsed.

The amount of rich metadata propagated at the end of the auditing process depends significantly on the quality and quantity of associations that the Auditing tool has been able to create between the crawled metadata and the RSS feeds’ metadata. We have achieved good results for only a small number of selected journals. The tool was not applied to all the journals because its software was still under development when the ticTOCs Project was completed. Being “metadata extraction” an inherently uncertain method and prone to propagate errors, the tool only augments metadata if a 100% matching between web-crawled and RSS feed metadata key values (e.g. article’s title) is obtained. The initial results have shown that the auditing tool effectiveness can be increased by allowing human intervention to resolve the cases when no 100% matching was achieved. This is in agreement with Liu’s findings [Liu, 2007] that not all metadata can be generated automatically. Applying fuzzy logic models to express or map the human’s semantic for the “matching” concepts into an exact numeric range of acceptance or rejection is an approach being explored to enhance the chances of identifying logic units from HTML document sections.

3.4. Data Mapping and Normalization Tool

This tool combines the principles of Berson’s third and fourth tools (Data parsing and standardization and, Data extraction, transformation and loading tools.) The tool first atomizes the TOC RSS item metadata elements in data units that are subsequently compared (mapped) against the “ticTOCs Best Practices” metadata set proposed and recommended to publishers by the ticTOCs Project [Rogers, 2008]. This “Best Practices” metadata format includes the standard RSS 1.0 modules dc, prism, content and syndication.

This tool also “normalizes” the cleansed and audited metadata in the sense that it produces uniform, consistent, valid and well-formed metadata values for each journal feed item and for all the harvested journal feeds. Thus the original metadata is converted to a canonical format which subsequently will be used by the service to expose TOC RSS feeds via APIs.

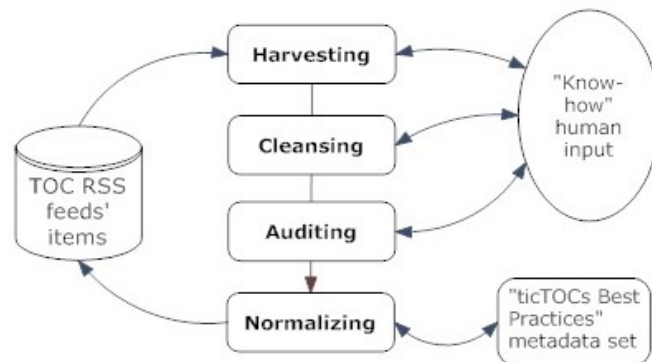


Figure 3: ticTOCron tools outline

4. Impact of ticTOCron

The first positive impact on a TOC RSS feeds service as a result of using ticTOCron is that its content is continuously being updated and consistently exposed to third parties. Additional impacts may become apparent as the usage of ticTOCs grows. The long term impact of our work will derive from the response it receives from TOC RSS feeds services, and its reuse by, or incorporation in, other related software systems. In the next section we explore possible applications that would be built using APIs to reuse the output generated by ticTOCron.

A clear benefit of using ticTOCron is that once a feed has been processed by ticTOCron, it will be available even when the publisher RSS site is temporarily unavailable or it is producing XML errors. In those cases the live feeds would not work, but the ticTOCs "cached" version will still be providing users with access to the online content of the TOC.

The second benefit is the richness of the metadata produced by ticTOCron. Most of the feeds provide titles of the articles as well as abstracts. However the abstracts are not always "abstracts." Additionally, although the title and abstract would be enough to let users quickly decide whether or not the article is relevant for them, they wouldn't be enough, for example, for creating citation lists. Thus, more metadata is required to create full identifications or citations (e.g. Authors, Vol, Issue, Date, etc.) By providing richer metadata, ticTOCron enables easy creation of bibliographic references and can indirectly make users more willing to visit the articles' website.

Providing uniform, consistent, valid and well-formed metadata values for all the journal feeds can also impact on the way in which other services or users could create persistent/stable URLs to articles from the TOC RSS feeds. As publishers tend to use different formats and structures for publishing the content of their articles, laborious work needs to be done by the user to create those URLs. For example, when a DOI has been included in the article abstract, the user would need to extract it manually to use in other applications such as reading lists. Enabling services such as openURL resolution from TOC RSS feed items is also made challenging by incomplete and inconsistent metadata. In short we believe that the work done with ticTOCron would increase the interoperability of TOC RSS feeds based services.

5. Possible Applications

Most of the possible applications of ticTOCron are connected with reusing what ticTOCron has produced to create multiple outputs, applications or services from that single source of journal TOCs content. The standard, consistent and rich metadata produced by ticTOCron allows service providers to single source journal TOC content in powerful ways. The potential of ticTOCron output would be increased by making small adjustments to its software or setup, such as keeping previous harvested feeds, as a consequence, past issues wouldn't be overwritten when a new issue is detected by the harvester (ticTOCs being a current awareness service, doesn't retain all items in any aggregated database. Thus, each time a new journal issue is published, items of the previous issue are removed.) To increase the reusability of ticTOCs, the service would need to retain items from "*n*" previous issues. The value of "*n*" is in function of the frequency in which the issues of a journal are published and could be used to calculate the number of issues that need to be retained for each journal.

Another feasible application would be to create a service or facility to generate a localised version of ticTOCs for an Institutional Library. The advantage of ticTOCs for this purpose would be that ticTOCs is likely to be more up to date than any other database, such as ZETOC (Electronic Table of Contents, <http://zetoc.mimas.ac.uk>), SCOPUS (<http://www.scopus.com>) or similar.

Once there is enough back-file data, a ticTOCron-based database would be used to produce information on "research tendencies", to answer queries such as "in which subject or discipline have authors been writing in the last six months" or "what are the recent trends in recent research in my area of interest."

An application in the field of Institutional Repositories (IRs) would be a facility to alert IR managers when a submitted paper from their IR has been published in a scholarly journal. IR managers need to know when the peer reviewed version of a paper is available. The fact that the TOC RSS metadata has been augmented by ticTOCron makes TOC RSS feeds a viable solution to help IR managers because it is expected that the number of matches between "submitted" versions (from the IR) and "published" versions (from ticTOCs) has a direct relationship with the quality of the metadata produced by ticTOCron.

6. Conclusion

This research indicates that generators using both extraction and harvesting methods have the potential to create useful metadata from TOC RSS feeds. Although the results obtained are preliminary as the ticTOCs service is new, the results show real possibilities of success and draws attention to important areas of metadata generation practice.

ticTOCron has been able to enhance the TOC RSS feeds for most of the 12,321 journals in ticTOCs. Although more work is needed in order to make further use of the Auditing Tool, ticTOCron has shown itself to be a potential tool to facilitate interoperability of TOC RSS feeds.

The process of enhancing TOC RSS feeds cannot be completely automated. As in any application of technology for generating metadata, the best results are achieved when humans are able to “teach” the software as part of a continuous iterative process. This conclusion is in agreement with what researchers have concluded that the most effective mean of metadata creation is to integrate both human and automatic methods [e.g., Schwartz, 2002; Craven, 2001]. As it is shown in Figure 3, human intervention occurs in all the stages of ticTOCs application. The result is still a sustainable and scalable computational model because ticTOCron uses “human intervention” to learn from “experience” (TOC RSS feeds have structural patterns that are unlikely to change frequently.) Consequently the needs for human intervention and computational resources should decrease with the time.

A specific conclusion from this study is that using cleansing, auditing, mapping and normalizing tools, derived from the Berson’s model, is a suitable approach for creating optimal TOC metadata and for augmenting the quality of harvested TOC RSS feeds.

The study also has confirmed that a correct use of RDF modules [5] can be a key ingredient for any successful service built on top of a database of TOC RSS feeds. The benefit of using those modules would be enhanced by following simple best practices for interoperability, such as do not restrict access to TOC RSS feeds, produce regular static copies of RSS feeds for fast processing and make available up-to-date OPML [6] files, etc.

Acknowledgements

ticTOCron R & D is based upon work supported by ticTOCs, a JISC funded project under a “Users & Innovation: Personalising Technologies” grant. The authors acknowledge the generous support of the entire ticTOCs team. The methodologies, findings and conclusions described in this manuscript are those of the authors and do not necessarily reflect the views of JISC or ticTOCs.

References

- Aeleen Frisch, A. (2002) Essential System Administration: Help for UNIX System Administrators. *Published by O'Reilly, ISBN 0596003439, 9780596003432 1149, Third Edition. pp. 90-99*
- Albassuny, B.M. (2008) Automatic metadata generation applications: a survey study. *Int. J. Metadata, Semantics and Ontologies, Vol. 3, No. 4, pp.260–282.*
- Berson Alex, Dubov Larry and Dubov Lawrence. (2007) Master Data Management and Customer Data Integration for a Global Enterprise. *Book published by McGraw-Hill Professional, ISBN 0072263490, 9780072263497. 406 pages*
- Craven, T., (2001) Changes in Metatag descriptions over time. *First Monday, Vol. 6, No. 10 - 1 Oct. 2001*
- Chumbe, S., MacLeod, R., Barker, P., Moffat, M. and Rist, R. (2006) Overcoming the obstacles of harvesting and searching digital repositories from federated searching toolkits, and embedding them in VLEs. *Proceedings 2nd International Conference on Computer Science and Information Systems, Athens, Greece.: <http://eprints.rclis.org/archive/00006394>*

- Farooq, U., Ganoë, Craig H., Carroll, John M., Council, Isaac G., and Giles, C. Lee (2008). Design and evaluation of awareness mechanisms in CiteSeer. *Information Processing and Management*, 44, pp. 596–612
- Greenberg, J. (2004). Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications. *Journal of Internet Cataloguing*, 6(4): pp. 59-82
- Greenberg, J., Spurgin, K. and Crystal, A. (2006). Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *Int. J. Metadata, Semantics and Ontologies*, Vol. 1, No. 1, pp.3–20.
- Hammersley, B (2005) Developing feeds with RSS and Atom. *Book published by O'Reilly, ISBN 0596008813, 9780596008819. 276 pages*
- Lagoze, C., Krafft, D., Cornwell, T., Dushay, N., Eckstrom, D., Saylor, J. (2006) Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience. *Proceedings of the 6th ACM/IEEE-CS joint Int. Conference on Digital Libraries*. pp. 230-239
- Liu, J. (2007) Metadata and its Applications in the Digital Library: Approaches and Practices. *Published by Libraries Unlimited, London, pp.143–149.*
- Manola, F. and Miller, E. (2004) RDF Primer. *W3C recommendation*.
<http://www.uazuay.edu.ec/bibliotecas/conectividad/pdf/RDF%20Primer.pdf>
- Margaritopoulos, M., Margaritopoulos, T., Kotini, I. and Manitsaris, A. (2008). Automatic metadata generation by utilising pre-existing metadata of related resources. *Int. J. Metadata, Semantics and Ontologies*, Vol. 3, No. 4, pp.292–304.
- Reuven M. Lerner, R., (2004) At the forge: aggregating syndication feeds. *Linux Journal, published by Specialized Systems Consultants, Inc., ISSN: 1075-3583. Issue 128 (December 2004), Page 7.*
- Rogers, L. (2008) RSS and scholarly journal tables of contents: the ticTOCs project, and good practice guidelines for publishers. *FUMSI Magazine, October 2008 [online] URL: <http://web.fumsi.com/go/article/share/3356>*
- Schwartz, C. (2002). Sorting out the web: approaches to subject access. Westport, Connecticut: Ablex publishing. *Part of the Contemporary Studies in Information Management, Policies, and Services series by Herson, P (Ed.)*
- Tonkin, E., Muller, H. (2008) Keyword and metadata extraction from pre-prints. *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008. pp. 30-44*
- Van de Sompel, H., and Lagoze, C. (2001) The Open Archives Initiative Protocol for Metadata Harvesting. *URL: http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html*
- Wittenbrink Heinz. (2005) RSS and Atom: Understanding And Implementing Content Feeds & Syndication. *Packt Publishing. ISBN 1904811574, 9781904811572. 250 pages*

Notes

- [1] "RSS feeds" is a mechanism to efficiently deliver frequently updated content to users (There is some discussion as to what RSS stands for, but most people plump for 'Really Simple Syndication'. In essence, the feeds themselves are just XML files, designed to be read by computers rather than people).
- [2] ticTOCs is a project funded by JISC to prototype a journal Tables Of Contents (TOCs) current awareness service, available at <http://www.tictocs.ac.uk> . ticTOCron has been developed as part of ticTOCs.
- [3] CrossRef is a not-for-profit association to enable easy identification and use of electronic content (<http://www.crossref.org>)
- [4] Document Object Identifier (<http://www.doi.org>)
- [5] RDF RSS 1.0 Specification and Modules. <http://web.resource.org/rss/1.0>
PRISM Module: http://www.prismstandard.org/resources/mod_prism.html
- [6] OPML (Outline Processor Markup Language): <http://validator.opml.org>