

Bibliographic metadata (including citation)

Emma Tonkin and Alexey Strelnikov
{e.tonkin, a.strelnikov}@ukoln.ac.uk

Introduction

Metadata extraction is the process of describing extrinsic and intrinsic qualities of the resource such as document, image, video, etc. As the result of that a number of records are produced which enables efficient search, sort and mining functionalities provided by the repository engine. For example extraction of the resource title would enable search and sort by title. Since the process of extracting is well defined it could be automated in order to speed up a resource upload to the repository and ensure the quality of extracted metadata.

The common or so called formal extraction includes collecting information what is already generated by resource creation tool and operating system along with the facts what could be gathered from the content itself. As notes Golub et al (2009), the formal metadata extraction could occur from the resource content analysis, like, for example document structure analysis or bibliographic citation analysis. The techniques of content analysis are non-deterministic, i.e. those give results with some probability, because of the nature of methods lying behind them.

Some formal metadata types are specific to the particular resource types, an illustration of that being the fact that 'page count' make sense for documents, but not for video files.

Type	Name	Example
Intrinsic/Formal	Filetype/attributes	PDF, MOV, MP3 at given bitrate and encoding
Intrinsic	File size	Size of file
Intrinsic	Resource language	e.g., Video contains audio streams in English, French and Russian.
Intrinsic	Type of document	Preprint, technical report, magazine article, journal article, MSc thesis, homework, PowerPoint presentation, poster
Intrinsic	Title	"A Christmas Carol"
Intrinsic	Author(s)	Charles Dickens
Intrinsic	Affiliation or contact details of author(s)	
Intrinsic	Date of publication	Year, (may include month, day, and time)
Intrinsic	Page count	
Intrinsic	Document index, table of contents	
Intrinsic	Sources cited/referenced within document/ bibliography	
Extrinsic	Theme	Poverty
Extrinsic	Related documents	Some forms of metadata explicitly encode various types of relationship between document objects

Table 1. Some examples of data types and corresponding formal metadata

Many approaches to metadata extraction are based on document structure (Greenberg, Spurgin, and Crystal, 2006). Document structure involves the use of the visual grammar of pages, for example,

making use of the observation that title, author(s) and affiliation(s) generally appear in content header information. At least five general structures may be instrumental in metadata extraction:

- **Formatting structure:** The document may have structure imposed on it in its electronic format. For example, from an HTML document one can extract a DOM tree, and find HTML tags such as <TITLE>.
- **Visual structure:** The document may have a prescribed visual structure. For example, postscript and PDF specify how text is to be laid out on a page, and this can be used to identify sections of the text.
- **Document layout:** The document may be structured following some tradition. For example, it may start with a title, then the authors, and end with a number of references.
- **Bibliographic citation analysis:** Documents that are interlinked via citation linking or co-authorship analysis may be analysed via bibliometric methods, making available various types of information.
- **Linguistic structure:** The document will have linguistic structure that may be accessible. For example, if the document is written in English, the authors may "conclude that xxx .", which gives some meaning to the words between the conclude and the full stop.

Metadata can be extracted via various means, for example using support vector machines upon linguistic features (Han et al., 2003), a variable hidden Markov model (Takasu, 2003), or a heuristic approach (Bergmark, 2000). Han et al. (2006) describe an approach that makes use of the following models upon formatting information: Perceptron with Uneven Margins, Maximum Entropy (ME), Maximum Entropy Markov Model (MEMM), Voted Perceptron Model (VP), and Conditional Random Fields (CRF). Various approaches that are useful in metadata extraction include:

- **Classification**, with the example of Bayesian classification
- **Pattern matching**, with examples of regular expressions
- **Direct application of observed heuristics**
- **Model fitting:** where prior knowledge is available, it may be applied as domain knowledge to build a set of models for use in metadata extraction.
- **Elicitation of grammatical structure** (ideally automated). This enables probabilistic parsing. For this, various approaches may be taken, including Hidden Markov Models, Maximum Entropy Markov Models and conditional random fields are also discussed.

The many methods available today have different uses, competences and areas of weakness so it is likely that a complete metadata extraction tool will make use of several approaches for different tasks.

Why is this important?

Automatic metadata generation has sometimes been posited as a solution to the 'metadata bottleneck' that repositories and portals are facing as they struggle to provide resource discovery metadata for a rapidly growing number of new digital resources (Polfreman et al, 2008). Automated metadata extraction would save time and efforts for both resource uploaders and repository managers. Tool support could fully or semi-automated, in other words would allow user to check and correct suggested values to more precise. In both cases tool support would prevent expensive manual creation and allow to expand number of collected metadata records. Manual generation is expensive per record – the generation of one record takes a significant amount of time, with estimates dependant on the record type; document deposit in an institutional repository takes somewhat over five minutes (Carr & Harnad, 2005), although with more complex application profiles the times increase significantly. Multiplying this by a huge number of electronic resources (mainly documents) means that the cost to the institution is significant, especially if experts must be employed to create or review this data. Manual generation is error-prone – the same is very much of

automated metadata extraction, but there is a significant difference between the types of error that are seen.

Several automated metadata extraction tools exist, both in terms of prototypes and openly available or commercial packages. However, integration between these tools and potential usage environments is limited, for several reasons; firstly, data extraction packages often need to be trained for a precise problem area – many tools are specifically built for use within a specific topic area, such as physics, computer science or the arts. Furthermore, the problem area often defines precisely which features of the document to extract. A system designed to support image search will attempt to extract images and their captions from PDF documents. An institutional repository manager will require the extraction of document features suitable for a Dublin Core application profile such as OAI-DC, or perhaps SWAP – which requires a far more complex series of document features, including relationships between draft and final versions of documents amongst other things. A formal metadata extraction system designed to support Learning Object Metadata (LOM) will require information about the document from the perspective of the learner as well as formal metadata such as object title and creator. Hence, there is a need to adapt existing software, as well as potentially making use of several software packages or services, in order to create a single complete record within a given use case and environment.

Some Scenarios

Title	Ensuring consistent metadata usage/style
Author	Alexey Strelnikov and Emma Tonkin
Narrative	Bob is a repository manager, and he is aware that there are many limitations to the metadata that is currently in the institutional repository that he administrates. He decides that it is necessary to improve the quality of the metadata. Usually, he would do this via a manual process of inspecting each record to ensure that metadata fields are correctly formed and formatted, and contain the correct sort of information - as well as looking for errors such as word substitution in titles or other 'sloppy' user errors. However, in this case he decides to use a new tool that has been added to his repository installation, which inspects both metadata and source material (in this case PDF/DOC/HTML formatted eprints) and detects omissions, departure from policy, and inconsistencies between sources of information. It provides a batch mode interface to suggest the records that are in gravest need of improvement and rank errors by frequency and severity, enabling him to solve problems in order of decreasing severity. In this way manual inspection of records becomes a less time-consuming and necessary process.

Title	Supporting the user in creating metadata records
Author	Alexey Strelnikov and Emma Tonkin
Narrative	Dominic is an end-user of a document repository, and has decided that he is

	<p>going to add a number of pieces of work that he has recently published to the repository. Usually this would mean that he has to complete several separate processes of uploading documents and typing in formal metadata about those documents. This includes author information such as email address, name and affiliation, title, category and subject - often present in ACM-formatted papers, as keywords are often present in other paper formats. It also includes references, especially since the repository that he is using is making use of the SWAP application profile, and stores lots of information taken from within the paper.</p> <p>However, on this occasion he decides to make use of the new interface that allows him to select all of the documents he wants to upload, and upload them all together. It then presents him with a series of records containing the information that he would normally have had to type in, highlighting any missing pieces of data so that he can fix them, and enabling him to make any corrections that may be required. Because the records are tailored to the document type there are very few extraneous pieces of information, and very few unnecessary fields; because the documents have already been uploaded and analysed the number of keystrokes required is greatly lessened and the problem of creating new metadata is transferred mostly into a problem of proofreading extracted records for accuracy.</p> <p>Although Dominic does not know it, the metadata presented to him is improved from the original information extracted from the paper by working from information extracted via reference parsing from other papers - this allows additional information to be collected and stored, and exposed to the user where necessary.</p>
--	---

Title	Using inter-repository consistency analysis to inform the AP review and redesign cycle
Author	Alexey Strelnikov and Emma Tonkin
Narrative	Edward is responsible for an application profile designed for use with scientific datasets. He would like to be able to report back to the committee who share responsibility for future developments in the standard, regarding the level of current use of the application profile, but also regarding the consistency and quality of records produced using that application profile. In other words, he wants to find out whether people in different institutions apply the application profile in a consistent manner. In order to find this out he looks at the data available from an aggregator, which uses OAI-PMH or OAI-ORE to harvest the records from within each system, and then compares the content of the records for consistency.

Title	Measuring impact of work within and outside a research community
-------	---

Author	Alexey Strelnikov and Emma Tonkin
Narrative	Fiona is a researcher looking to create a final report for a project lasting several years. She decides that she wants to demonstrate that her work during this project has had a significant impact within her research community. In order to do this, she looks up her paper on several well-known services such as citeseer.net and Google Scholar. This allows her to see how many citations her various papers and articles have received. She is then able to describe the impact of her work in terms of inspiring other work, and its reception within the general research area. She can also see where and for whom her work had an impact, and which aspect of the work was most interesting to others.

Each scenario should make clear whether it is based on: a) existing technologies which are available to any university, b) experimental technologies which are currently being evaluated; c) potential which exists but is not yet being developed.

Tools

The process of automated metadata extraction is not one that can generally be satisfied by a single piece of software. This is because there are several stages in the process; a typical model for extraction of features present within the text itself is that of first extracting text, and then running the generated text stream through an appropriate set of tools in order to identify the information present within it. There are exceptions to this – for example, in some cases such as HTML or XML mark-up it is sometimes possible to extract information from named tags within the document, and of course in general many formats including PDF provide existing approaches to embedding document metadata, of which one may also wish to take advantage. Techniques have also been described that directly make use of formatting information within the document to attempt to identify document features. However, it is generally appropriate to make use of as many approaches as possible (a case of 'the more the merrier').

Hence, one focus of interest in building a robust tool-chain for the extraction of metadata from a large and heterogeneous document set is the identification of a large number of import filters. These may range from tools such as pdftotxt and antiword to APIs such as that provided by OpenOffice. A greater number of filters will simply confer the ability to try the same eventual feature identification mechanisms on a larger number of formats. The following tools mostly make use of a series of different text extraction tools. It is also worth noting that optical character recognition is often useful for this purpose as well, since it confers the ability to produce a text stream from a PDF or other format encoded as an image. Once the text is extracted or made available, it is then used to identify features.

Here are some examples of tools that have been applied within the JISC information environment in the past. There are many other tools, libraries and prototypes that perform functions of relevance to the extraction of formal metadata.

- DC-Dot – a UKOLN tool from 2000 that extracts metadata from HTML documents, as DC metadata; the tool was rewritten in early 2009 to add input/output plugins for further formats, such as Microsoft Word and PowerPoint, and to improve the range and number of formats handled.

- DepositPlait – a tool developed by Aberystwyth that extracts metadata from several document formats for use in repository metadata creation
- DataFountains – generates metadata from HTML pages if given a URL, described by Polfreman (2008) following the MetaTools evaluation as giving the best output of the tools that they had tested.
- paperBase – a prototype tool developed at the University of Bristol (Tonkin and Muller, 2008), that extracts metadata from document text. Designed and trained for use on scientific papers, Polfreman (2008) remarked that 'output from all of the tools was considered to be disappointing and markedly inferior to the quality of metadata that Tonkin and Muller report that paperBase has extracted from scholarly works'.

Also note that many metadata extraction tools will also make use of existing data sources in order to improve their analysis. This is generally handled by the application itself.

Standards

There are many standards of relevance to this process. We have already covered the fact that the document import process is dependent on the existence of an appropriate import filter allowing the document to be read, and have briefly mentioned at the beginning of this document that it is also important to consider the metadata standard(s) that are to be applied as outputs, in deciding which methods and services are appropriate to underpin the work. In this section we briefly identify some standards that are of interest, and identify the phase during the extraction process at which they come into play.

Document encodings

Various current and historical document formats are currently widely supported in metadata extraction, whilst there are others that are clearly of interest but enjoy only patchy or limited support from the tools mentioned here. In particular, PostScript (PS) and Portable Document Format (PDF) formats are very well supported. HTML is supported, but the wide variety of potential uses and the display/style role played by HTML encoding can make this more difficult to extract reliably – the document template can be an issue. OpenXML and OpenOffice's formats are important, but support for them is very limited. Legacy Microsoft formats such as DOC and PPT enjoy a limited level of support at present in the tools described here.

An additional encoding issue that has a significant impact on document text extraction, and hence on metadata extraction, is that of font and encoding. If these are not known at the time of document ingest, or if the document was encoded in a non-standard manner or with certain methods of partial font embedding, then the document will not be readable or fully or correctly readable (eg. the result will be partially or entirely corrupt). Furthermore, some documents are encrypted or the use of text extraction has been intentionally limited in some manner; an example of this is the use of the PDF usage limiting or encryption systems.

Metadata encoding standards

Several encoding standards are of relevance to us – in particular, output support for Dublin Core (OAI-DC, Scholarly Works Application Profile...) is important in order to integrate services with the institutional repository landscape. Additionally, OAI-ORE (Open Archives Initiative - Object Reuse and Exchange) is increasingly relevant to the JISC community for the description and exchange of aggregations of Web resources (Open Archives Initiative, 2009). Different communities will of course make use of a variety of metadata standards, but the vast majority of the tools mentioned above presently support only a subset of DC. RDF is also important in this domain, but may be used to encode both Dublin Core and ORE amongst other possibilities;

therefore it can be seen as an overarching superset of this listing.

Feature encodings

Much of the work that is done in feature extraction comes down to an understanding of the formatting that denotes certain entities. For example, the ability to correctly identify that a substring represents a date or an email address is a significant part of the overall task of classifying document features. This is usually done by means of targeted regular expressions. The greater the adherence to standards in the creation of a document, the easier it typically is to extract this information in this manner. However, even a seemingly simple task such as the extraction of dates from a document can be a complex problem, since there are many potential irregularities resulting from either the use of non-standard encodings, of widely seen encodings that are subject to ambiguities (such as the use of day/month/year vs month/day/year), or of local conventions or simply of localized forms that are not covered by the regular expression. In particular information in languages other than English can suffer from this problem, since regular expressions are generally limited to the language of the author.

Citation and reference formats

A reference contains the bibliographic information for a paper or other type of resource, such as an article or web page, and is designed in order to enable the reader to look up the resource. A citation is an inline label that identifies the reference within the context of the document itself. For example, *Delaney, 2009 describe...*

There are many different formatting standards for bibliographic references and for citations, such as for example the APA format, used in psychology, education, and other social sciences; the MLA format, used in literature, arts, and humanities; the AMA format, applied in medicine, health, and biological sciences; the Turabian standard, designed for use in any subject area; the Chicago standard, used by books, magazines, newspapers and so on (Delaney, 2009), not to mention Harvard, Numeric and MHRA (Thornes, 2008). However, this is by no means an exhaustive list; indeed, the number of reference formats actually in use is sufficiently large that it has been described as excessive and, as a result, remarkably expensive (Leslie & Hamilton, 2007).

The existence of multiple citation and reference styles has as an effect the fact that citation and reference extraction and analysis are rendered a more complex task than might otherwise be expected; not only are these often used in an inconsistent and idiosyncratic manner, but they are also used in a manner that is often peculiar to specific organizations or conferences. Hence there may be a need to make use of some amount of 'guesswork' in correctly parsing citations from a given document, a problem that is reminiscent of the more general issue of identifying features in a text string. However in general this is not as complex a problem as the more general issue of metadata extraction, and therefore it is possible to achieve a high level of accuracy across a document corpus.

Key Issues

It is absolutely necessary before text can be extracted from a document that (some proportion of...) the full-text resource is made available. This may mean providing the full document, or the front pages and the last pages – however, that would limit the use of that resource for certain purposes, such as ascertaining the original length of the document, and therefore it is strongly recommended that the full-text resource be made available wherever it is possible.

The readability of the full-text resources is a significant factor in the success of this process; as mentioned above, some document formats are in general much easier in terms of text extraction than others. As a result it is important for those hoping to make use of this sort of technology to consider this at all stages of their document deposit workflow. The format in which a document is published may have a significant impact on many forms of accessibility, and one of these forms is that of machine/automated reuse. However, if a significant number of resources exist in a format

that does not already have good support, it is often possible to create new filters and subsequently to reuse existing software for the rest of the process – an advantage of a process that is effectively modular in nature.

There are legal issues associated to the reuse or analysis of document full-text in order to extract information from it. Initially there is the problem of accessing the document at all; this may or may not be acceptable depending on the terms of service of the website or repository. There is the problem of publishing what is in effect a work derived from the document full-text (eg. The document surrogate, or metadata record). This is compounded by the fact that in some cases it is desirable to retain the original full-text of the document internally, in order to allow the reparsing of the document and subsequent analysis of its content, which may yield a better result as the system is retrained, gains further information about the area (through citation parsing, for example), and is able to make a better 'guess'.

One problem with the use of metadata to publish metadata records extracted, especially via statistical methods, from full-text documents, is the fact that the records are essentially works in progress. In practice of course the same objection can be made to metadata records created by human beings, since these also contain errors and require validation, editing by an expert and perhaps also further adaptation for use in a given area. However, with an automated method we typically have access to an estimate of how accurate the result that we have gained may be. This information (a confidence rating) can be retained and published – however, there is presently no widely accepted manner of publishing any form of confidence rating along with metadata elements or values. This means that this information is lost, meaning that client systems are unable to use that information in consuming the metadata.

Finally, language and localization is a significant problem in metadata extraction. Many of these tools work only in languages for which they are explicitly trained, particularly content-based machine learning techniques. However, others (such as those which depend on formatting information to detect which piece of information is which) may not be affected by this factor.

Recommendations

The availability of full-text documents is a key issue in enabling the use of this sort of mechanism. The legal and social aspects of this approach to metadata extraction are, however, potentially significant. The consequences of making this sort of metadata available should be explored.

Effective reuse of extracted information requires new conventions for publishing the confidence information along with the metadata. This requirement should be explored, along with the more general issue of indicating provenance and quality/consistency of metadata.

The production of a robust system that is easy to retrain for given areas of use, provides easily customizable output, and is able to deal with many types of document format, language and localization, is an important factor in practical adoption. Working towards this end should be a primary aim for the JISC.

Our final recommendation is to look more deeply into the problem of analysing citations in their context of use. For example, relative order and placement of the citation may allow the inference of relationships between the papers that are cited. Papers cited together, for example, may be expected to have some level of similarity. It is also useful to classify citations according to their nature – such as technical or background citations.

Acknowledgements

The authors thank Henk Muller for his feedback.

References

- Bergmark, D. (2000). Automatic Extraction of Reference Linking Information from Online Documents. CSTR 2000-1821, Cornell Digital Library Research Group.
- Carr, L. and Harnad, S. (2005). Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving. Retrieved from <http://eprints.ecs.soton.ac.uk/10688/>.
- Delaney, R. (2009). Citation Style for Research Papers. Retrieved June 30th from <http://www.liu.edu/cwis/cwp/library/workshop/citation.htm>
- Golub, K., Muller, H., Tonkin E. (2009). *Technologies for metadata extraction*.
- 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.
- Han, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D. and Zheng, Q. (2006). Automatic extraction of titles from general documents using machine learning. *Information Processing and Management*, Volume 42, Issue 5, September 2006, Pages 1276-1293.
- Han, H., Giles, C. L., Manavoglu, E. and Zha, H. (2003). Automatic Document Metadata Extraction using Support Vector Machines, *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM Press, New York, pp.37-48
- Han, H., Giles, C. L., Zha, H., Li, C., and Tsioutsoulouklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries*, ACM Press, New York. pp. 296-30
- Han, H., Zha, H., and Giles, C. L. (2005). Name disambiguation in author citations using a K-way spectral clustering method. In *Proceedings of JCDL'2005*. pp.334-343
- Leslie, D. M. and Hamilton, M. J. (2007). Editorial: A Plea for a Common Citation Format in Scientific Serials. *Serials Review*, Volume 33, Number 1, 2007.
- Open Archives Initiative (2009). Open Archives Initiative - Object Reuse and Exchange. Retrieved on July 2nd 2009 from <http://www.openarchives.org/ore/>
- Polfireman, M. (2008). *Metatools final report*. JISC
- Takasu, A. (2003) 'Bibliographic attribute extraction from erroneous references based on a statistical model', *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries*, ACM Press, New York, pp.49-60.
- Thornes, S. (2008). References and Citations Explained. Retrieved on 3rd July 2009 from http://library.leeds.ac.uk/info/200201/training/218/references_and_citations_explained
- Tonkin, E. and Muller, H. L. (2008). Semi Automated Metadata Extraction for Preprints Archives. *JCDL 2008*.