Jeremy L. McLaughlin
July 2015

Turning hamburgers into a cow: an introductory comparison of PDF metadata extraction using two reference management systems

Reference management systems have been around for decades. Software like EndNote and web-based platforms like RefWorks provide researchers with a variety of functionality around collecting, managing, and citing sources for academic, scholarly, and general purposes. New platforms like Mendeley, Zotero, and ProQuest Flow (now RefWorks)* mix collaboration and sharing functionality with advanced website scraping and Portable Document Format (PDF) metadata harvesting capabilities. These technologies are designed to enhance the user's workflow around discovery and collection of digital resources. While several studies have compared the ability of these tools to extract PDF metadata within the larger context of PDF metadata extraction, these tools, like many others, fall victim to the problems associated with PDF metadata in general.

As more content, from more sources, is made available online, the study of metadata extraction from standard formats of academic content is a significant area of research in library and information science and computer science (see, for example, Lipinski, Yao, Breitinger, Beel, & Gipp (2013) for a comparison of tools using two common approaches to automated metadata extraction). Much of this research focuses on PDF metadata as catalogers rely on as much automation as possible to secure data that describes the main features of a work and, in turn, information seekers rely on this data for discovery and retrieval (Marinai, 2009; Lipinski, Yao, Breitinger, Beel, & Gipp, 2013).

Despite the need for more metadata, and the continued and growing use of PDF as a distribution format a disconnect remains between content publishers and content users. In reviewing the growing chasm between scholarly publishing and the needs of readers, Pettifer et al., (2011) notes that even though PDF is the format of choice for an overwhelming majority of downloaded scholarly content, attempts to extract meaning from PDFs as they are currently used is similar to trying to turn a hamburger back into a cow.

The purpose of this study is to compare the bibliographic metadata harvesting capabilities of two reference management systems: Mendeley and ProQuest Flow. Secondary research questions are related to the availability of XMP metadata in published scholarly material, and whether this has any impact on metadata extraction. The following sections of this paper will examine the literature related to attempts to extract PDF metadata using automated harvesting technologies and the role of reference management systems in PDF organization for individual researchers. The next section details the research methodology, data analysis, and results, followed by a discussion of the research findings.

The results of this study speak to the capabilities of individual researchers, without the benefit of large-scale extraction tools, to properly harvest metadata for their own purposes using platforms like Mendeley and ProQuest Flow. Additionally, extraction results and the performance of RMS technology in this study hint at their enhancement and possible large-scale use in digital library environments. Despite the hopeful possibilities, concerns remain about the format of published scholarly material in PDF form that make standardization of metadata extraction a continued challenge.

*The metadata extraction analyses for the study were completed before ProQuest Flow became RefWorks

PDF Metadata and XMP

Metadata is traditionally thought of as 'data about data' or 'information about information' but, by definition, metadata needs to be structured and is supposed to define or add descriptive layers to the data it accompanies/that accompanies it. This critical aspect of metadata supports information extraction, storage, discovery, access, and retrieval.

The Portable Document Format (PDF) was created by Adobe in 1993 to aid in the distribution of digital documents providing view-and-print anywhere capabilities. In 2008 Adobe released PDF as open standard. While access to, and use of, PDF has grown exponentially over the last two decades, their use as a format for distribution and dissemination in scholarly communication has shed light on the problem of interoperability. Specifically, issues related to discovery and automation around sources of unstructured text (Adefowoke Ojokoh, Sunday Adewale, & Oluwole Falaki, 2009). Individual journal articles (born digital or PDF versions of print articles) are made of up text, images, tables, graphs, links, data, citations, etc… By default, these structural entities are not individually searchable, even if the parts can be of stand-alone value to researchers (e.g. a specific equation or charts).

As an ever-increasing amount of content is made available on the Web, and as more and more of this content is available in PDF, the accurate harvesting of structured data about published works has become tantamount to libraries and scholars. The PDFlib Whitepaper on metadata support in PDFs (found at http://www.pdflib.com/fileadmin/pdflib/pdf/whitepaper/Whitepaper-XMP-metadata-in-PDFlib-products.pdf) provides the following examples of common PDF metadata properties:

- The author of a PDF document.
- The date a PDF document was created or a JPEG image was taken with a camera.
- The name of the photographer who took an image.
- The serial number of a personalized document.
- The year of manufacture of the engineering product described in a document.
- The reference number of a document in a legal case.

While PDF has become the de facto standard in scholarly publishing critical issues and concerns of metadata quality associated with the common PDF have remained unexplored (Bui & Park, 2013). PDF metadata can be stored using the Adobe Extensible Markup Platform (XMP) which was released in Adobe Acrobat 5 in 2001. XMP is an XLS-based format that allows users to label and embed metadata defining properties for document and image characteristics into various file formats, including PDF, TIFF, and JPEG. The XMP specification travels with the PDF and includes more than a dozen predefined schemas, the most widely used – according to the PDFlib website (http://www.pdflib.com/knowledge-base/xmp-metadata/) – being Dublin Core. Since early 2012, XMP is also an International Standards Organization (ISO) standard (16684-1).

The Adobe website (http://www.adobe.com/products/xmp.html) lists the following key benefits of XMP:

- Create smart assets that retain their context when traveling across software, devices, and databases.
- Provide full extensibility by adding arbitrary metadata to media while visualizing it in Adobe products.
- Enable powerful search and retrieval of rich media across diverse file formats and database systems.

- Manage relationships of assets throughout their lifecycle of content creation and consumption.
- Build on open standards and open source licenses to foster a common exchange across the industry

Soon after its release, Rosenblatt (2002) noted that the goal of XMP was to make metadata accessible and standardized (sets of elements and what they describe). However, early on it was noted that XMP does not solve all problems…namely, consistency in values that speak to the overall need for controlled vocabularies and categorization technologies. Roszkiewicz (2005) points to an overall lack of support by Adobe as an early reason for resistance to XMP. He goes on to note that while the XMP situation is likely to improve as the product matures, users will continue to be frustrated by how manual the process is.

XMP's future potential improved significantly in 2009 when the journal *Nature* announced that it would begin using XMP to embed complete metadata for all articles from the December 18, 2008 issue forward. "In a collaborative project with Nature Publishing Group, Charlesworth's development team created a tailored workflow solution to handle the work. The resulting bespoke system of automated batch-processing extracts information from XML files, and produces standardized validated PDF files with the XMP metadata embedded" (Nature Publishing Group, 2009, p. 1).

While XMP is a formal ISO Standard, the use of individual record or batched PDF metadata remains a significant area of concern. As a result of the publisher focus on PDF as a tool for reading and printing, varying styles used by publisher, and the need for author contributions for much of the content deposited into open and institutional repositories (especially for pre-print versions, PDFs created directly by individual researchers, and gray literature) in most cases articles contain little to no descriptive or administrative XMP metadata. As a result, publishers and readers of PDFs have lost – and continue to lose – the opportunity to advance scholarly communication in significant ways (Willinsky, Garnett, & Wong, 2012).

While Petifer et al. (2011) calls for a more "dynamic format" than the PDF in general, Marinai (2009) points out that collecting and managing even the main descriptive features of a work (author, title, publication date) is costly and time consuming. To maximize their extensive reach, PDF should be used for reading, for printing, for sharing and mobility, for linking to the web and/or other related documents, and for their metadata (whether linked through XMP or extracted through other means). In both cases, when it comes to the most widely used source of distributed PDF content – the academic journal article – most of the opportunities, and burdens, for proper metadata creation fall to the publisher at the point of publication (Willinsky, Garnett, & Wong, 2012).

PDF Metadata Extraction

To help make the best of the current situation many studies are focused on extracting valuable data from PDFs without the existence of pre-defined bibliographic metadata. The need for metadata revolves around access and use of PDF for numerous reasons associated with bibliographic data (descriptive or administrative metadata). Metadata is essential for digital library functions, for large scale mining, for research assessment purposes, and for individual researchers who harvest, read, and cite these articles in their research. Most current research focuses on identification and extraction of metadata to create new or enhanced records that aid in archiving, discovery, analysis, or for other institutional use.

Working with freely available products, bulk processing, and machine readable outputs, Lipinski, Yao, Breitinger, Beel, & Gipp, (2013) compared seven tools that use two commonly identifiable approaches across many projects in this area: structure/content analysis (heuristics – document

placement and styles of font), and machine learning (SVM – support vector machines; HMM – hidden Markov models; CRF – conditional random fields). This study focused on extraction of title, author full name, author last name, abstract, and year of publication and found that the GROBID CRF implementation performed best, followed by Mendeley Desktop SVM and SciPlore Extract content analysis.

Similarly, Hsiao, Chang, & Thomas (2014) propose an HMM model to extract PDF titles and authors that were then enhanced with cross-referencing against online digital libraries. Adefowoke Ojokoh, Sunday Adewale, and Oluwole Falaki, (2009) designed a model for extraction based on segmentation of articles by the hierarchy of the division of the document (spacing, style, keywords) and rule-based pattern matching using PHP, Java, MySQL, and HTML.

Granitzer, Hristakeva, Jack, and Knight (2012) and Granitzer, Hristakeva, Knight, Jack, and Kern (2012) use a heuristic-based approach to compare extraction techniques of various products and found that the 2-stage SVM used by Mendeley was very successful in "scraping" additional metadata from PDF sources and helped establish the benefits of bibliographic metadata management in a crowdsourced environment. Mendeley takes advantage of an extensive online database to enhance metadata. This is a common approach – thought not perfect – given the large-level data inconsistencies in XMP packets and from extraction methods and typically relies on bibliographic databases to retrieve additional metadata based on digital identifiers or specific fields (author and title, for example) (Aumüller, 2009; Marinai, 2009; Hsiao, Chang, & Thomas, 2014).

The key takeaway from these studies is that while they bring about varying results, metadata quality and consistent use with PDFs is the root of the problem. Research in large-scale metadata extraction from PDF is interesting but these studies have a limited focus (first page extraction; focus on one piece of metadata; metadata for/from a specific discipline or data source) which makes it hard to compare or standardize results (Lipinski, Yao, Breitinger, Beel, & Gipp, 2013). Building tools with extraction in mind helps discovery and improves overall availability of metadata (Adefowoke Ojokoh, Sunday Adewale, & Oluwole Falaki, 2009) for digital libraries. The larger issue of resource discovery and sharing highlight the growing awareness for automation but these studies do not speak to the day-to-day issues faced by researchers as they interact with scholarly content in PDF.

Reference management, PDF metadata, and the individual researcher

Individuals need metadata for their own research purposes and, whether they know it or not, use PDF metadata with reference management tools as they manage personal collections of research literature. This is an ongoing process of search and discovery which can result in individual researchers harvesting thousands of PDFs throughout their career.

By using reference management tools to store and manage PDFs, individuals are creating their own personal digital libraries (Hull, Pettifer, & Kell, 2008). Despite this, very few examples in the literature focus on PDF metadata discovery at the level of the individual researcher. Aumüller (2009) examines the problem of discoverability and searchability resulting from incorrect metadata on individual researcher's personal computers. Extracting metadata from scholarly articles is difficult for individual researchers, but as the Aumüller study points out, manually establishing proper mapping back to web resources (by matching local PDFs to a metadata repository, in this case Google Scholar) is effective for managing PDFs and extracting additional metadata from them.

What the literature does present are comparisons of reference management system functionality, including the performance of tools like RefWorks, Mendeley, and Zotero in their ability to collect and manage research materials.

Lorenzetti – survey of 78 researchers found that 79.5% had used a reference management tool and that 98% has used EndNote, Reference Manager, or RefWorks.

They are primarily used (Fitz, Francese).

Basak limited. Zhang extensive but no PDF details.

Mead & Berryman. Gilmoour. PDF usage in reference management. "new twist" of PDF management. Studies comparing reference management tools. Focus on importing and accuracy of references.

Individual scholars can use Mendeley or Zotero to manage PDFs and extract metadata to varying degrees of success. Within these social platforms or in crowdsourced metadata resources user-driven enhancements can make a significant difference in the quality of extracted metadata (Granitzer, Hristakeva, Jack, & Knight, 2012). Despite their success and ease of use, reference management tools are not always the most successful performers when compared to other methods for automated metadata harvesting, even with enhanced records from digital libraries (Marinai, 2009; Lipinski, Yao, Breitinger, Beel, & Gipp, 2013; Hsiao, Chang, & Thomas, 2014).

But unless they specifically enhance the metadata of the PDFs they encounter, or use or design a more advanced tool for their personal research workflow, the use of PDF metadata by and for individuals highlights the problems of incorrect or altogether missing metadata as the norm instead of the alternative. (Willinsky, Garnett, & Wong, 2012).

Given the need for extraction methods and the importance placed therein, there is surprisingly little comparative research on metadata quality or the use of PDF XMP by certain publishers or in specific journal content (e.g. metadata in publisher versions of scholarly articles, not versions created by authors and uploaded to repositories).

METHODOLOGY

This study seeks to compare the performance of two reference management systems (Mendeley and ProQuest Flow) in their ability to extract bibliographic metadata from scholarly PDFs. In doing so, we also analyze the presence of bibliographic XMP metadata in PDFs from various publishers in various years, and whether this has any effect on platform performance. The research questions under consideration are supported by the literature, and include:

- how do Mendeley and Flow perform in extracting PDF metadata (with or without the existence of PDF XMP metadata)?
- does Mendeley present superior performance in extracting PDF metadata overall?
- what is the ratio of harvested PDFs with XMP metadata versus those without?
- which of the core bibliographic fields (author, title, journal name, date of publication) are most available in PDF XMP metadata?
- how often does PDF XMP metadata refer back to a URI or other external resource?
- is there a noticeable difference in PDF XMP metadata availability from *Nature* articles pre- and post-2009?
- how does the PDF XMP metadata in *Nature* compare to other sources of scholarly articles?
- are theses a reliable source of PDF XMP metadata?

Data collection was carried out towards the goal of assessing the availability of PDF XMP metadata and comparing the PDF metadata extraction in the two reference management platforms. PDFs were harvested from the San Jose State University Martin Luther King Jr. Library using Nature Online, ProQuest Theses and Dissertations, and the SFX link resolver. Each PDF was given a numerical file name, and to assist in fully answering the proposed research questions, several articles were manually given XMP metadata while – for additional comparative purposes – others were duplicated and the duplicates were manually provided XMP metadata. For this initial study, 30 articles were used, including 22 from Nature (14 pre-2009 (6 with enhanced PDF XMP metadata) and 8 post-2009) and 8 theses from ProQuest.

Full bibliographic citations for all articles used were collected from the databases along with the PDF, or were manually created from the PDF when file naming convention/article number coding took place. Each PDF was initially reviewed for the visibility – in any location – of XMP metadata for: author name (author of the scholarly work, not author of the PDF), title of work, journal or publication title, date of publication (not date of PDF creation), and the use of a URI in the record. Next, each PDF was uploaded to the reference management platforms via drag-n-drop and individual records were reviewed for the proper extraction of author name, title of work, publication title, date of publication, and the use of a URI or other source link that could be used to enhance metadata matching.

Based on a scoring method developed from Lipinski, Yao, Breitinger, Beel, & Gipp, (2013) each article was scored based on the existence and accuracy of PDF XMP metadata and the performance of each reference management platform in extracting and populating the appropriate record field with accurate PDF metadata. Scoring used the following rules: 0 = no metadata; .5 = incorrect metadata (did not apply to formatting concerns e.g. capitalized journal titles or author naming conventions); 1 = correct metadata.

Results and discussion

This analysis yielded some very interesting results. Across the entire dataset (n=30) PDF XMP metadata was available in 26 PDFs but only 41% of the fields reviewed (author name, title of work, journal or publication title, date of publication, and use of a URI) were available directly in the PDF XMP. All of the PDFs from articles published in 2010 (4) had complete XMP metadata. None of the PDFs from 1997 had any metadata and PDFs from 2000 only included 8% of the fields in the XMP metadata.

Nature articles made up a majority of this initial test dataset and results verified the inclusion of PDF XMP metadata in all Nature PDFs after 2009. For all nature PDFs across all years (n=22 including 6 modified) just over 56% of the metadata fields reviewed were present in the PDF XMP. Post-2009 *Nature* PDFs also include multiple URIs leading back to various data sources.

Theses did not perform well with 0 having any XMP metadata available.

	total PDF	total XMP	XMP poss	Flow total	Flow poss	Men total	Men poss	XMP Result	Flow result	Men result
ALL 30	30	62	150	95.5	150	108	150	41.33%	63.67%	71.67%
1997	4	0	20	9.5	20	10	20	0%	48%	50%
1997 mod	4	13	20	11	20	10	20	65%	55%	50%
2000	8	3	40	24.5	40	31	40	8%	61%	78%
2000 mod	2	6	10	9	10	10	10	60%	90%	100%
2010	4	20	20	20	20	20	20	100%	100%	100%
2014	8	20	40	21.5	40	26.5	40	50%	54%	66%
nature	22	62	110	59.5	110	70	110	56.36%	54.09%	63.64%
1997	4	0	20	9.5	20	10	20	0%	48%	50%
1997 mod	4	13	20	11	20	10	20	65%	55%	50%
2000	4	3	20	16.5	20	20	20	15%	83%	100%
2000 mod	2	6	10	9	10	10	10	60%	90%	100%
2010	4	20	20	20	20	20	20	100%	100%	100%
2014	4	20	20	13.5	20	20	20	100%	68%	100%
modified	6	19	30	20	30	20	30	63%	67%	67%
theses	8	0	40	16	40	17.5	40	0.00%	40.00%	43.75%
2000	4	0	20	8	20	11	20	0%	40%	55%
2014	4	0	20	8	20	6.5	20	0%	40%	33%

In the PDF XMP metadata (with 41% of fields reviewed available across the dataset) URI (8/30) and publication date (9/30) scored the lowest. Title scored highest with 16/30. The modified records helped these scores for 1997 and 2000 (and overall). While the 1997 un-modified Nature articles included no XMP metadata, those from 2000 did have some authors and titles (though primarily incorrect). Additionally, it is important to highlight that the 2010 and 2014 Nature records did include 100% of PDF XMP metadata while theses from 2000 and from 2014 included 0%.

With very few exceptions, Mendeley outperformed Flow in extracting PDF metadata regardless of overall XMP metadata availability. Mendeley extracted 8% more of the fields reviewed and successfully extracted 100% of metadata fields from PDFs from articles published in Nature in 2000, 2010, and 2014. As noted in the literature, Mendeley offer bibliographic data enhancement which provides additional metadata from the Mendeley catalog. On PDF upload users are given the option to review the details and mark them as correct or to search the catalog. This latter feature was attempted with several PDFs in the dataset but none retrieved any additional bibliographic metadata from the Mendeley catalog.

Flow scored highest at retrieving PDF titles (23.5/30) and URI (20.5/30). It is important to note that the high score for Flow related to URI is because Flow automatically adds the URL for “retrieved from” which includes the link back to the database the PDF was retrieved from. Flow retrieved the least amount of journal titles (only scoring 11.5/30).

Modified records from 1997 and 2000 (n=6) had 63% of metadata fields and Mendeley and Flow both extracted 67% of fields from these PDFs. The details of these results show an interesting inconsistency in the improvement in reference management metadata harvesting results. This is especially true for PDFs from 2000 which saw a drastic improvement. However, while modified 1997 PDF XMP metadata resulted in increased harvesting compared to unmodified PDFs, the results of both Flow and Mendeley show retrieval of less fields than were available. These results require additional analysis using a larger dataset of modified and unmodified records for comparison.

CONCLUSIONS

This study helps to illuminate the problems associated with PDF metadata extraction and use, at the individual and the institutional level. The research findings further validate that PDF XMP metadata associated with scholarly articles are not consistently accessible or standardized, and that XMP metadata availability may vary widely within an individual publisher's PDFs, and in general from publisher-to-publisher, journal-to-journal, and PDF-to-PDF.

Previous research into the metadata harvesting capabilities of reference management tools were further verified, with Mendeley producing the best results when compared to ProQuest Flow.

Nature set the bar in a noticeable way in terms of PDF XMP metadata availability for the fields studied. This was highlighted by the use and extraction – by Mendeley – of multiple URIs from post-2009 Nature PDFs. However, a significant difference between PDF XMP metadata and the extraction results in both tools questions the usefulness of XMP metadata in general. Additional study will further explore the concerns related to PDF XMP modification at the user level and its impact – or lack thereof – on reference management tool harvesting. Initial testing would suggest that the focus on mostly large-scale, technical methods for extracting proper metadata from PDFs is more promising than "fixing" PDF XMP. This serves a much larger purpose but also builds off of the productivity associated with newer generation reference management tools.

To increase statistical relevance and the depth of comparative analysis, additional testing will include a larger set of theses as well as a more general set of PDFs from various publishers across multiple years, and with additional modified PDF XMP fields. Zotero, another reference management tool with PDF metadata extraction capabilities, will also be included in the discussion.

If it is available, XMP metadata certainly could be used by individual researchers or en-masse for digital libraries in an automated way…but this is almost too little and definitely too late. In some ways it is almost ironic that end-users have access to enhanced metadata extraction based on SVM in Mendeley and other tools while library administrators struggle to find appropriate technology to index and catalog growing digital collections. To maximize the use of scholarly PDF metadata across platforms and users, it is imperative that ILS and catalogues, and vendors of ebooks and other digital content, find ways to utilize PDF metadata extraction technologies similar to those being used by reference management tools. Identifying the best method is just part of the solution; publishers should consider standardizing the format for digital or PDF versions of journal articles.

Until the best extraction method can be developed and agreed on, or until existing commercial methods can be maximized for open library use, a majority of the useful metadata found in the holy cow that is the scholarly PDF will remain as minimally structured hamburgers.

References

Adefowoke Ojokoh, B., Sunday Adewale, O., & Oluwole Falaki, S. (2009). Automated document metadata extraction. *Journal of Information Science, 35*(5), 563-570.

Aumüller, D. (2009). *Retrieving metadata for your local scholarly papers.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.8872&rep=rep1&type=pdf#page=595

Bui, Y., & Park, J. (2013). An assessment of metadata quality: A case study of the national science digital library metadata repository. Paper presented at the *Annual Conference of CAIS/Congrès Annuel De L'ACSI.*

Granitzer, M., Hristakeva, M., Jack, K., & Knight, R. (2012). A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management. Paper presented at the *27th Annual ACM Symposium on Applied Computing*, 962–964.

Granitzer, M., Hristakeva, M., Knight, R., Jack, K., & Kern, R. (2012). A comparison of layout based bibliographic metadata extraction techniques. Paper presented at the *2nd International Conference on Web Intelligence, Mining and Semantics.*

Hsiao, W., Chang, T., & Thomas, E. (2014). Extracting bibliographical data for PDF documents with HMM and external resources. *Program: Electronic Library and Information Systems, 48*(3), 293-313.

Hull, D., Pettifer, S. R., & Kell, D. B. (2008). Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Computational Biology, 4*(10).

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. Paper presented at the *13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 385–386.

Marinai, S. (2009). Metadata extraction from PDF papers for digital library ingest. Paper presented at the *10th International Conference on Document Analysis and Recognition*,

Nature Publishing Group. (2009). *Charlesworth helps nature publishing group semantically enrich PDFs*. http://www.nature.com.libaccess.sjlibrary.org/press_releases/charlesworth.html: Nature Publishing Group.

Pettifer, McDermott, Marsh, Thorne, Villeger, & Attwood. (2011). Ceci n'est pas un hamburger: Modelling and representing the scholarly article. *Learned Publishing, 24*(3), 207-220.

Rosenblatt, B. (2002). XMP: The path to metadata salvation? *Seybold Report: Analyzing Publishing Technologies, 2*, 3.

Roszkiewicz, R. (2005). The brief, tortured life of XMP. *Seybold Report: Analyzing Publishing Technologies, 5*, 19.

Willinsky, J., Garnett, A., & Pan Wong, A. (2012). Refurbishing the camelot of scholarship: How to improve the digital contribution of the PDF research article. *Journal of Electronic Publishing, 15*(1).