VRA Bulletin

Volume 45
Issue 2 *Cataloging Today: Enlarging the Sphere*

Article 4

March 2019

Metadata Automation: The Current Landscape and Future Developments

Marlee Graser
Southern Illinois University Edwardsville, magrase@siue.edu

Melissa Burel
Alabama A&M University, melissa.burel@aamu.edu

Metadata Automation: The Current Landscape and Future Developments

Abstract
As a profession, librarians are already seeing the ways in which automation is challenging traditional methods of cataloging and raising questions about the future of manual cataloging work. Workflows for metadata creation—from the most basic to those involving data scraping, harvesting from APIs, and data migration and manipulation—indicate a shift from a cataloger's traditional role of metadata creation to technologically-oriented metadata harvesting and management. Additional new technologies, like deep learning computation, are beginning to address the call for automated metadata creation for visual resources, reinforcing this shift and creating new opportunities for innovative workflows and description. New technologies have the potential to profoundly impact the ways that libraries ready themselves and their data for the semantic-web environment and redefine cataloging work moving forward. Will these further automation advances really change the role of the metadata librarian? This article argues that the continued escalation of automation and linked data in the semantic web will only be a continuation of metadata librarians' current technological skills and commitment to data quality control.

Keywords
automation, metadata, metadata workflows, library profession, Automatic Image Annotation (AIA)

Author Bio & Acknowledgements
Marlee Graser received her MSLIS from the University of Illinois-Champaign-Urbana in 2014. She is currently serving as the Metadata Librarian at Southern Illinois University in Edwardsville where she enjoys researching ways to enhance the library's role as a nexus of learning and develop the library as a service center for students, faculty, and the community.

Melissa Burel is the Metadata Cataloging Librarian at Alabama A&M University. She received her MLIS in 2012 from Wayne State University and her MM in 2008 from Western Michigan University. Her professional interests include user-needs assessment, data analysis, and the organization of information.

This feature articles is available in VRA Bulletin: https://online.vraweb.org/vrab/vol45/iss2/4

Introduction

As a profession, metadata librarianship is rapidly evolving with the creation of new technical innovations. Within the past few decades, libraries and cultural heritage institutions are increasingly answering the call to expand access to their rare and unique materials through digitization and online publication. Combined with the boom in born-digital materials, library professionals have had to retool and rethink the methods through which they perform their duties. A large part of this shift has been the rise of technologies that make it possible to automate large portions of the work or to use automated techniques to drastically cut down on the time or effort involved in creating, migrating, and repurposing metadata. The methods of automation that metadata librarians utilize within their workflows are as diverse as the collection formats, metadata schemas, and digital collection platforms available and automation can occur in every step of the process from creating an XML schema, to migrating one metadata schema to another, to extracting metadata for other uses. There are also new possibilities in metadata automation coming out of machine learning and computer vision research. This article is a selected review reporting on the current use of metadata automation in the field, an investigation into the possible applications of machine learning for automation in the future, as well as a discussion of automation's impact on the library metadata professional.

Automation for XML Schema Creation

Often in the description of digital objects, information professionals guide student workers to fill out Excel spreadsheets with appropriate metadata. On some platforms, such as ContentDM, these spreadsheets, saved as CSV files and formatted with Dublin Core headers, can then be directly ingested to describe a digital object or digital collection. Other platforms, such as Islandora, require metadata in MODS XML format. While some institutions have developed the technology for spreadsheet ingestion, most institutions must convert metadata in Excel documents to XML.

In his presentation, “Finding a New M.O.: Metadata Automation on a Budget at a Medium-Sized Institution” Nicholson (2016) described his workflows to process Excel metadata in bulk and convert them to MODS. Nicholson’s workflow began with uploading Excel documents containing the metadata for roughly 100,000 images into OpenRefine. OpenRefine, previously known as GoogleRefine, is a free online tool that is useful in the management and manipulation of data. Within OpenRefine, Nicholson recommended cleaning the data by using clustering, faceting, and the Google Refine Expression language. He also walked through the process of separating out multiple subject headings and their corresponding URIs. The next step involved exporting the metadata to a

MODS template, which can be copied and pasted from a text editor, such as Wordpad. Once he completed the export, the resulting file was a large batch of MODS records that required some further editing to remove null values, extra quotation marks, and the separating out of individual records utilizing XSLT. This workflow is advertised by Nicholson as being relatively straight forward and requires the information professional to know minimal scripting or coding.

Python is another tool that information professionals can use to create XML data. Python is a free, object-oriented scripting language that is used in the fields of data analysis and data manipulation. It has modules explicitly built for working with XML data and is often considered one of the more natural programming languages to learn and execute, skipping compiling steps common in other languages. Another positive aspect of Python is that it integrates well with other languages, opening up more possibilities in application creation. Should a librarian want to adopt the use of Python within a metadata workflow, there are many freely available guides, books, and classes to learn Python scripting (Lutz and Ascher 1999).

Bartczak and Glendon (2017) described a process that utilized Python to build their XML records in their article “Python, Google Sheets, and the Thesaurus for Graphic Materials for Efficient Metadata Project Workflows.” The library had to digitize over 100,000 photographs in time for the University of Virginia’s bicentennial anniversary. Their workflow began with digitizing the items while students entered basic descriptive metadata, usually transcribed from the item or the folder from which it originated. Once these items were in the system, the Metadata Analysis and Design Team read it into Python (essentially they loaded the data into Python) and utilized Python’s Panda module to massage the data into a 2-dimensional array (rows and columns), assigning new headings, removing unnecessary information, and splitting field information. The team employed an automated method of error checking as well as some manual corrections before the data was converted to a CSV file and downloaded into Google Sheets. In Google Sheets, students entered metadata specifically focused on title, description, and subject headings, followed by peer review and a final review by a metadata librarian. Once the Google Sheets were deemed ready, the librarian used a script in the Python lxml module to build the MODS XML records. Reading the CSV file in as a data frame, the Python script created the header information and then iterated over each element in the data frame to build the XML. Bartczak’s and Glendon’s Python code is available on GitHub for further reference.

Automation for Schema Conversion and Metadata Repurposing

Automation is also frequently used as a method to reuse and convert metadata into various schema to expedite the description of large collections or migrate existing metadata to a new platform with different schema requirements. An automation tool that is used to convert XML into different schemas is XSLT. XSLT (Extensible Stylesheet Language Transformations) is a programming language that is used to transform the content and structure of XML documents (Kay 2008). XSLT templates consist of rules that when applied to XML create entirely new XML outputs. Along with template rules, it also relies on pattern matching, which is implemented using XPath expressions. One of the many useful features of XSLT is its looping ability, or recursion. The XSLT can loop through an element that contains multiple pieces of information (for instance a subject heading and its URI) and parse out the pieces and assign them to different elements. Within the library field, there are many examples of XSLT that have been used successfully that an information professional can easily copy and adjust to fit a specific need (Cole, et al. 2018, 58-67).

Averkamp and Lee (2009) described their workflow for converting Proquest UMI Dissertation Publishing metadata into Dublin Core (DC) using an XSLT style sheet for use in their bepress institutional repository in, “Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository.” Their process involved taking the Proquest batch of XML files and utilizing Microsoft Office products to compile these into one document. The team then applied XSLT to crosswalk the data from Proquest’s schema to Dublin Core. While the XSLT style sheet did the work of manipulating the different fields, information professionals had to determine the correct field mappings to create valid DC XML. They also created unique URLs to each ETD, normalized metadata, and performed a manual review. The use of automation allowed the ETDs to be available to the public sooner than if each one was manually cataloged. These new records in bepress facilitated the next step of transformation into MARC records for the local catalog. While the transition from DC to MARC may not supply perfect records, it does expedite large portions of the resource description for the local catalog.

XSLT is a powerful tool for transforming metadata and customizing display but is useful only if the data is harvested and readily available. Many digital asset and content management systems include plugins or feeds that allow for the automated harvesting of metadata using the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH). This feature enables library professionals to harvest metadata from any repository that supports the protocol to repurpose records for value-added services, such as discovery layers or other metadata aggregators (Lagoze, et al. 2015). OAI-PMH standardizes the set of rules that define how systems communicate with one another to request or respond to requests for shareable metadata. Responses always take place in the XML syntax, and while OAI-PMH can use any metadata encoded in XML, it always supports the unqualified Dublin Core schema, so there is a minimum common agreement (Kapidakis, et al. 2015, 1).

In their presentation, "Pipe Dreams: Harvesting Local Collections in Primo Using OAI-PMH," Rinna, et al. (2018), of Western Michigan University, discuss their workflow for harvesting records from ContentDM, bepress, Luna Insight, LibGuides, and ArchivesSpace via OAI-PMH into their discovery layer, Primo. The process required the configuration of data sources to ensure that the metadata pulled from each source repository would map directly into the schema used by Primo. After configuring the metadata in their data sources, they verified that the XML documents were valid by using an OAI-PMH validator. Within Primo, they configured the data sources, scope value, normalization rule set, and mapping tables to import the data in pipes. After troubleshooting each data source's unique issues, Western Michigan University was able to successfully load records of digital objects, finding aids, scholarly publications, and LibGuides into their discovery layer.

While OAI-PMH provides a gateway to the data stored in various platforms, often the data retrieved by OAI-PMH needs to be substantially manipulated for display purposes or ingestion into different platforms. Multiple case studies demonstrate how libraries have leveraged OAI-PMH, along with XSLT transformations, to improve user experience, build new services, and automate the repurposing of existing metadata. Librarians at Pennsylvania State University Libraries used OAI-PMH feeds to automate the creation of catalog records from the metadata supplied by authors when submitting their ETDs (Robinson, et al. 2016). OAI-PMH was essential in the process because it provided the data that the librarians could later manipulate into RDA-compliant MARC records. However, substantial customizations occurred after the harvest through the use of the tool MarcEdit, an application developed by Terry Reese that is used for the creation and manipulation of metadata in various forms. Within MarcEdit, information professionals used the unqualified Dublin Core metadata harvested by the OAI-PMH feed and processed it through an XSLT crosswalk that transformed the metadata. This process required editing the XSLT transformation they were using to map author-supplied Dublin Core metadata to the appropriate RDA-compliant MARC fields. They also had to consider how to include MARC fields that lacked author-provided metadata, such as fixed-length data fields. This required further edits to the customized XSLT transformation. While these customizations were time-consuming, Robinson, et al. found that harvesting metadata through the repository's OAI-PMH feeds and manipulating the metadata using MarcEdit and customized XSLT transformations saved substantial amounts of time, stating, “the time required to process a semester’s worth of ETDs plummeted from 100-160 hours to fewer than 8 hours” (2016, 195).

Similarly, in “Why Purchase When You Can Repurpose? Using Crosswalks to Enhance User Access,” Keenan (2010) discussed a project to convert approximately 168,800 records that describe the resources in the U.S. Congressional Serial Set database. When given the option to purchase the MARC records or receive free Dublin Core records, Keenan and her team opted to repurpose the Dublin Core metadata for use in their local catalog. The team harvested the records over OAI-PMH with MarcEdit and utilized XSLT to crosswalk the data from Dublin Core to MARC, which took into consideration the local ILS’s indexing capabilities. They also used MarcEdit for mass field editing and metadata normalization. Keenan described challenges that were encountered with batch loading into the local ILS, demonstrating some of the unforeseen stumbling blocks when working with different systems. While this workflow description and the XSLT the author provided is useful for any library that needs to accomplish a similar task, the budget comparison for purchasing new records versus repurposing existing metadata is very compelling. While there were challenges associated with creating a new workflow to convert metadata, in dollars and cents the library saved roughly $24,500 with metadata repurposing.

Automation for Metadata Extraction and Enhancement

There are cases in which librarians can enhance their metadata during the extraction and conversion process. Examples of this can be seen in rather straightforward ways, such as Robinson, et al. (2016) or Veve (2016) utilizing XSLT to add in RDA fields while transforming their XML records. In Glerum and Bortmas’s (2015) presentation regarding the conversion of bepress ETD metadata to MODS, they also outlined a process to enhance their metadata by extracting information from the ETD PDFs. After using XSLT to convert Proquest metadata to MODS, the duo used javascript to remove the bepress title pages of the ETDs and collected the text from pages one and two of the dissertations. The scraped data provided more description information, and XSLT and XProc were used to add the data into the MODS XML. XProc is a processing language that programmers use to string together multiple XSLT XML so that it processes chains of XML transformations (Kay 2008). While human review was necessary, this method created ‘thick metadata’ to enhance description while utilizing a semi-automated approach that saved considerable time and effort.

Randtke’s (2013) article, “Automated Metadata Creation: Possibilities and Pitfalls,” described utilizing digitization and automation to create a database for the Florida Administrative Code (FC). Because this publication continually changes with the addition of new rules and the removal of outdated information, the project presented unique challenges. Randtke and her team wanted to capture each update, the date they occurred, as well as digitizing the item to increase accessibility. Her team's workflow began with scanning each page of the code as a PDF, utilizing Adobe Acrobat to run optical character recognition (OCR), and using a PDF to Excel Extractor to import the text into Excel. Once in Excel, her team conducted a comparison of the Excel text to the item itself to determine the rules of how each piece of information would be extracted from Excel and mapped to each metadata field. A programmer translated these rules into the Visual Basic scripting language, and Randtke and her team ran the script, which organized the information into an Excel spreadsheet. After some manual cleanup, the team imported the final Excel data into the database. While errors did occur, the team utilized manual cleanup and review in multiple steps. Overall Randtke found that the computer error rate was low and accomplished the bulk of the work.

APIs also allow library professionals to reuse existing content in new ways to offer new or better services. An API, as defined by the Digital Public Library of America (DPLA), "is made up of a set of defined methods that someone can use to communicate with a (frequently complex) software system and get back responses in a way that a computer (and, with some practice, a human) can understand. A request is a URL sent to the web server over HTTP with the expectation of getting resource items back in the form of human-readable text or data" (API Basics n.d.). In this way, an API is able to return results from a request and allow libraries to reuse content from disparate sites. APIs function similarly to OAI-PMH in that it provides a gateway to content that might otherwise exist in a silo. Bullen (2016), currently of the Illinois State Library, utilized the ContentDM API to customize a website based on the content available through his institution's ContentDM. He accomplished this by using the API to query and retrieve data and then used PHP and Perl scripts to customize the resulting response. Bullen maintains a WordPress site, "A Cookbook of Methods for Using ContentDM APIs" that details his process and gives samples of his code. He provides further documentation on the project through GitHub.

Similarly, Gordon (2018) used an API with other tools to create a clip library of digitized audiovisual content. Since staff had already utilized ArchivesSpace to house the content description at the item level, the goal was to automate the output of a spreadsheet that connected metadata to a digital file. To do this, Gordon wrote a Python script that made use of the command line tool FFmpeg to automate the production of A/V clips. The Python script also utilized the ArchivesSpace API to create a spreadsheet of descriptive metadata based on the original A/V file. The filename and refID derived from the original file in ArchivesSpace connected the metadata and clip on the spreadsheet, which resulted in a ‘lightweight' searchable clip library that made finding and using A/V clips in-house faster and easier.

Machine Learning for Image Metadata Creation

Traditionally, libraries and cultural heritage institutions have made their materials available through a metadata creation and publication process in which metadata is created and applied to resources manually. Metadata created in this traditional way requires a vast amount of time and expertise to employ. Even if institutions can take advantage of the various methods to automate metadata conversion, repurposing, and extraction, the majority of descriptions themselves are still being applied ‘by hand.’

While there are methods by which some administrative, technical, and descriptive metadata is produced as digital images are created—either during the digitization process or at the time a born-digital image is produced—this initial process often only creates minimal metadata for images. More robust metadata that drastically increase the relevancy and efficiency of indexing and searching must still be created manually. This is a problem not just for cultural heritage institutions, but for individuals and corporations alike, as the number of images and visual resources being created through the use of readily available technologies has seen a meteoric rise within the past few decades. Optical character recognition programs have streamlined the way that textual data, like transcriptions, are applied and substantially decreased the amount of time spent creating descriptions for these resources. This technology allows for enhanced metadata that would have otherwise been too time-consuming to create. In the same way, computer scientists and information professionals have begun to look for ways in which the creation and application of metadata for visual resources can be automated to both save time and enhance the user experience.

In response to this need, a field of research has developed over the past three decades to address the time-consuming process of metadata creation for visual resources. Automatic image annotation (AIA), as defined by Cheng, et al. is “concerned with models/algorithms to label images by their semantic contents or to explore the similarity between image features and semantic contents with high efficiency and low subjectivity. Relevant labels are predicted for untagged images from a label” (2018, 242). As surveyed by Bhagat and Choudary (2018), the landscape of AIA has shifted as new research emerges. A complete survey on the state of AIA research is beyond the scope of this article, but, initially, AIA focused on content-based image retrieval, through which algorithms indexed low-level image contents through image processing to group data into object silhouettes, clusters of points, and/or image features (Smeulders, et al. 2000, 1356-1357). This allows for search and retrieval by matching patterns, object-recognition, and through the process of similarity (1373).

Within content-based image retrieval, the semantic gap is a significant issue. This gap is the disparity between the low-level content in images and the high-level semantic concepts that they might represent to a user. It is a relevant consideration for retrieval based on content because, as Datta, et al. point out, visual similarity and semantic similarity do not always mean the same thing (2008, 5:2). For example, if a user is performing a search based only on an image, rather than a text-based query, the “content-based image retrieval has to be conducted only in the visual feature space, but the performance is evaluated in the textual feature space” (Wang, et al. 2008, 355). While two images may contain the same color or texture as the query image, the meaning of the images retrieved based on this similarity may not be relevant to the user. In the second wave of AIA research, researchers have set out to address the semantic gap more efficiently by “finding the correlation between visual and textual features” relying heavily on machine learning (Bhagat and Choudary 2018, 3). Machine-learning essentially trains computers to learn the correlation between image features and textual data from examples given to it of annotated images (Jin, et al. 2004, 892). Programs then use the correlations that they’ve ‘learned’ to predict and apply textual data, like keywords, given the presence of specific image features (892). Generally, the more and better the data given to the machine, the more accurate its correlations will be. This marked a transition towards a focus on text-based image retrieval, or, more accurately, text-based image retrieval using the lessons learned from content-based image retrieval in which image queries are tied to semantic concepts through AIA and text-based queries are correlated to features within an image.

As stated above, the purpose of automatic image annotation is to streamline the process of image retrieval by automating the application of semantic meaning to image features. However, nearly all of the models for AIA discussed above base their predictive correlations on a training dataset of manually annotated images. While it may seem counterintuitive for the success of machine learning to rely so heavily on human input, the concept of ground-truthing is essential when ‘teaching’ a machine the meaning of image features. Ground-truth images are used to determine how accurately the machine ‘learned’ the semantic meaning of an image and is used as a metric for machine learning’s success. The concept of ground-truthing is also closely tied to the selection and development of an image set’s vocabulary. In nearly all of the AIA models described above, the programs depend on a dataset of ground-truth images annotated with a common, and often simple, vocabulary. There are examples of datasets, such as Iconclass, that have fairly well-developed vocabularies used in AIA (Hanbury 2008). It is in such cases that we can begin to see how machine-learning and AIA might extend to encompass the concepts of the semantic web easily into its algorithmic workflows, such as if the keywords that the machine applies are based on a vocabulary of triples and URIs.

A new wave of AIA research is emerging to address the time-consuming process of producing ground-truth images and manually annotated training image sets. In recent years, the research has focused on the promises of deep-learning techniques, most commonly using convolutional neural network-based features which mimic, as closely as possible, the process of human vision and image processing in a machine (Mayhew, et al. 2016). This uses multiple layers and interprets images as complex, multi-dimensional objects to interpret data within the image and correlate the image with its semantic meaning. The most recent research has turned toward deep-learning models that are semi- or unsupervised and models that "explore unsupervised image annotation techniques…where [the] training dataset is not labeled at all, only metadata (URL, surrounding texts, filename, etc.) are provided with a training dataset" (Bhagat and Choudary 2018, 3). This type of research is still in its early stages but is a promising direction for automating the description of large image sets using little to no human supervision and input.

Implications for the Metadata Librarian

A brief survey of the workflows summarized here indicate that metadata librarians are being called on to have knowledge and expertise not only in metadata description, application, and management in multiple schemas, but also in XML, open source tools for data cleaning, XSLT, Python, OAI-PMH, MarcEdit, and a host of other technical skills that are required to successfully manage metadata automation. Certainly, the evolution of metadata librarianship has begun to shift strongly toward computer science and programming skills rather than the more traditional skills generally associated with positions in cataloging and semantic description. Working in a team environment can help to alleviate the skills load required for one librarian. In the article "Establishing Sustainable Workflows for Cataloging and Metadata Services," Han states that over time, metadata work has become more collaborative and asserts that "because new standards and best practices are developed in many different areas of bibliographic control, it is impossible to expect one person in the unit to have all the knowledge and expertise" (2016, 310). Han goes on to state that while each member of a metadata team needs to work together to keep current on evolving metadata and linked-data standards, librarians should also network between units, subject specialists, and the IT departments at their institutions. It is important to note that a wealth of information can be shared throughout the library community as well. While there is a large variety of metadata schemas, digital collection platforms, and data types, small steps from different workflows can be customized to meet unmet needs.

Another application of automation that metadata librarians need to be involved in is the implementation of Linked Open Data (LOD) in digital collections. In the article “A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies,” Southwick (2015) described an exploratory project at University of Nevada, Las Vegas to implement LOD in their digital collections. The benefits of incorporating LOD have been written about extensively in the literature. It’s clear that LOD facilitates linking to the semantic web, breaks up silos within the library itself, is machine readable, and allows users more access to research materials, along with better search features. The literature, however, has very few case studies of institutions that have initiated this cutting-edge work. In her article, Southwick discussed the concepts related to Linked Open Data and demonstrated how their collections are incorporating these concepts in concrete steps; from using and creating URIs, selecting and following a data model, creating triples, and storing the RDF files in a publicly-available server. Southwick’s article provided numerous areas that an information professional can learn more, including the use of URIs, SPARQL, the creation of triples, the selection of data models, as well as some extended features available in OpenRefine. It is refreshing to see these concepts implemented in an LOD project. While the article is exploratory, it highlights a need in automation to address the demands of converting and creating Linked Open Data. Metadata librarians are in a unique position to foster this innovation and to prepare library data for the semantic web.

While automation is useful for working with large amounts of data in a timely and efficient manner, it is nothing without the metadata librarian’s knowhow. For example, in OpenRefine Nicholson (2016) recommends the use of faceting and clustering for metadata clean up. This process involves the gathering of faceted terms that allows a metadata librarian to identify inconsistencies in the data, including misspellings, irregular spacing, the use of both plural and singular forms of words, as well as inconsistent case use. Southwick (2016) also uses OpenRefine for reconciling subject headings against an LOD vocabulary, which required the analysis and selection of term matches. Additionally, Southwick describes how librarians had to apply their expertise to track and normalize local vocabularies that could not be reconciled against an LOD vocabulary, generate and store triples, and create and maintain URIs. In the workflow employed by Bartczak and Glendon (2017), descriptions and subject heading application were still carried out manually. Similarly, Averkamp and Lee (2009) found that, although the bulk of metadata transformation could be automated, the process required the expertise of a cataloger to apply topical subject headings and correct titles that included mis-capitalized acronyms after their conversion into MARC. When Rinna, et al. (2018) developed their workflow to build pipes to ingest metadata from multiple sources into Primo, their expertise was required to create normalization rules and address mapping issues, particularly when moving from the complex metadata in LUNA to simple Dublin Core used by the discovery layer. In the same way, librarians at Pennsylvania State University Libraries using OAI-PMH feeds to create catalog records from author-supplied ETD metadata needed a firm understanding of the standards set by RDA, MARC format, and Dublin Core to crosswalk the metadata and add fields that were missing from the source (Robinson, et al. 2016). In every article describing a workflow that involved automation, human quality control and discernment was required.

While automatic image annotation (AIA) has the potential to profoundly impact how metadata librarians do their work, the literature currently lacks examples of libraries adopting these models to apply metadata to their visual resources. This absence could be because the research is still largely in its developmental stages or because the type of metadata that librarians apply to the visual resources in their care is complex, robust, and contains details that are difficult to embed in AIA models. For example, the difference between photorealistic artwork and a photograph is difficult to discern at first glance even by a human. It is even more difficult to ‘explain’ to an algorithm within machine learning when the only information given to it are very similar annotated test images that have been ground-truthed with appropriate metadata. And because AIA is only as accurate as the ground-truthed test sets that are fed to it, it is unclear, at present, whether the technology will ever be able to discern nuanced semantic meaning in the same way that an expert metadata librarian does. A large part of a metadata librarian’s work is defining context, interpretation, and relationships that are absent from the pixels in the image itself. AIA might be able to describe the colors, objects, and even recognize specific people if given enough annotated examples from which it can match. However, its ability to tie abstract human concepts, like oppression, appropriation, racism, sexism, affection, or grief, to the image run through its algorithm seems beyond the capabilities of this specific technology because of the myriad of ways that these concepts are expressed in visual resources. It would be nearly impossible to build a ground-truthed test set that covered enough of the different ways these concepts are represented within visual resources that it could then begin to extrapolate these concepts to new images, particularly to more abstract works.

Automation provides a great opportunity for the library professional to create large amounts of robust metadata with qualities of linked open data to aid in the accessibility of library collections. While using automated metadata workflows has many positive aspects, such as expedited processing, greater data uniformity, and lower costs of production, it is also the librarian's job to mitigate some of the deficiencies associated with this type of metadata work. For example, information professionals will always need to conduct due diligence on the reliability of metadata sources and the completeness of the metadata description (Dobreva, et al. 2013). As cultural heritage institutions focus on digital access to their collections and methods of automation allow information professionals to repurpose metadata in new environments or extract metadata from different sources, the quality of metadata and the context that it describes become even more important. In this way, it is heartening to know that there are some processes within metadata librarianship that cannot be automated. Automation simply provides a mechanism through which information professionals can free up time to ensure that the focus remains on quality while expanding the capacity for quantity.

Bibliography

Averkamp, Shawn, and Joanna Lee. 2009. "Repurposing ProQuest Metadata for Batch Ingesting ETDs into an Institutional Repository." *Code4Lib Journal* 7. Retrieved from https://journal.code4lib.org/articles/1647.

Bartczak, Jeremy, and Ivey Glendon. 2017. "Python, Google Sheets, and the Thesaurus for Graphic Materials for Efficient Metadata Project Workflows." *Code4Lib Journal* 35. Retrieved from https://journal.code4lib.org/articles/12182.

Bhagat, P.K., and P. Choudary. 2018. "Image annotation: Then and now." *Image and Vision Computing* 80: 1-23. doi:10.1016/j.imavis.2018.09.017.

Bullen, Andrew. 2016. Using CONTENTdm's API to Customize Your Site and Access Your Data. December 7, 2016. Retrieved from https://www.oclc.org/developer/news/2016/using-contentdm-apis-to-customize-your-site-and-access-your-data.en.html.

Cheng, Qimin, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. 2018. "A survey and analysis on automatic image annotation." *Pattern Recognition* 79: 242-259. doi:10.1016/j.patcog.2018.02.2017.

Cole, Timothy W., Myung-Ja (MJ) K. Hahn, and Christine Schwartz. *Coding with XML for Efficiencies in Cataloging and Metadata: Practical Applications of XSD, XSLT, and XQuery*. Chicago: ALA Editions, 2018.

Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. 2008. "Image retrieval: Ideas, influences, and trends of the new age." *ACM Computing Surveys* 40 (2): 1-60. doi:10.1145/1348246.1348248.

Digital Public Library of America. n.d. API Basics. Retrieved from https://pro.dp.la/developers/api-basics/.

Dobreva, Milena, Yunhyong Kim, and Seamus Ross. 2013. "Instalment on 'Automated Metadata Generation.'" In DDC Digital Curation Reference Manual (1-35). Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/dcc_amg_final.pdf

Glerum, Annie, and Dominique Bortmas. "Migrating ETDs from Dublin Core to MODS: Automated Processes for Metadata Enhancement." Presented at a preconference for the American Library Association Annual Meeting, Orlando, FL, June 7, 2016. Retrieved from https://www.youtube.com/watch?v=VtxNAZvr3qM&feature=youtu.be.

Gordon, Bonnie. 2018. "Using Python, FFMPEG, and the ArchivesSpace API to Create a Lightweight Clip Library." *bloggERS! The Blog of the SAA's Electronic Records Section*. November 13, 2018. Retrieved from https://saaers.wordpress.com/2018/10/30/using-python-ffmpeg-and-the-archivesspace-api-to-create-a-lightweight-clip-library/.

Han, Myung-Ja K. 2016. "Establishing sustainable and scalable workflows for cataloging and metadata services." *Library Management* 37 (6/7):308-316. doi: 10.1108/LM-04-2016-0031.

Hanbury, Allen. 2008. "A survey of methods for image annotation*." Journal of Visual Language & Computing* 617-627. doi:10.1016/j.jvlc.2008.01.002.

Jin, Rong, Joyce Y. Chai, and Luo Si. 2004. "Effective automatic image annotation via a coherent language model and active learning." *MULTIMEDIA '04*. New York: ACM. 892-899. doi:10.1145/1027527.1027732.

Kapidakis, Santos, Nikos Houssos, Kostas Stamatis, and Panagiotis Koutsourakis. 2015. "Flexible Metadata Mapping Using OAI-PMH." PETRA '15 Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments. Corfu, Greece: ACM. 1-2. doi:10.1145/2769493.2769531.

Kay, Michael. *XSLT 2.0 and XPath 2.0 Programmer's Reference*. Hoboken: John Wiley & Sons, Incorporated, 2008.

Keenan, Teressa. 2010. "Why Purchase when you can Repurpose? Using Crosswalks to Enhance user Access." *Code4Lib Journal* 11. Retrieved from https://journal.code4lib.org/articles/3604.

Lagoze, Carl, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. "The Open Archives Initiative Protocol for Metadata Harvesting: Protocol Version 2.0." Last modified January 8, 2015. https://openarchives.org/OAI/2.0/openarchivesprotocol.htm.

Lutz, Mark, and David Ascher. *Learning Python*. Beijing: O'Reilly Media, 1999.

Mayhew, Michael B., Barry Chen, and Karl S. Ni. 2016. "Assessing semantic information in convolutional neural network representations of images via image annotation." 2016 IEEE International Conference on Image Processing (ICIP). 2266-2270. doi:10.1109/ICIP.2016.7532762.

Nicholson, Joseph R. "Finding a New M.O.: Metadata Automation on a Budget at a Medium-sized Institution." Presented at a preconference for the American Library Association Annual Meeting, Orlando, FL, June 7, 2016. Retrieved from https://www.youtube.com/watch?v=VtxNAZvr3qM&feature=youtu.be.

Randtke, Wilhelmina. 2013. "Automated Metadata Creation: Possibilities and Pitfalls." *Serials Librarian* 64 (1-4):267-284. doi: 10.1080/0361526X.2013.760286.

Rinna, Geraldine, Marianna Swierenga, and Emily Gross. "Pipe Dreams: Harvesting Local Collections in Primo Using OAI-PMH." Presented at Ex Libris Users of North America (ELUNA) 2018, May 3, 2018. Retrieved from: https://scholarworks.wmich.edu/library_presentations/15/.

Robinson, Ken, Jeff Edmunds, and Stephen C. Mattes. 2016. "Leveraging Author-Supplied Metadata, OAI-PMH, and XSLT to Catalog ETDs: A Case Study at a Large Research Library." *Library Resources & Technical Services* 60 (3): 191-203. https://elib.uah.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=117332860&site=ehost-live&scope=site.

Smeulders, Arnold W.M., Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. "Content-Based Image Retrieval at the End of the Early Years." *IEEE Transactions On Pattern Analysis and Machine Intelligence* 22 (12): 1349-1380. doi:10.1109/34.895972.

Southwick, Silva B. 2015. “A Guide for Transforming Digital Collections Metadata into Linked Data Using Open Source Technologies.” *Journal of Library Metadata* 15 (1): 1-35. doi:10.1080/19386389.2015.1007009.

Veve, Marielle. 2016. “From Digital Commons to OCLC: A Tailored Approach for Harvesting and Transforming ETD Metadata into High-Quality Records.” *Code4Lib Journal* 33. Retrieved from https://journal.code4lib.org/articles/11676.

Wang, Changhu, Lei Zhang, and Hong-Jiang Zhang. 2008. "Learning to reduce the semantic gap in web image retrieval and annotation." ACM SIGIR '08. Singapore: ACM. 355-362. doi:10.1145/1390334.1390396.