JESSICA CRAIG

MLIS Portfolio

University of California, Los Angeles
Department of Information Studies
Advisor: Miriam Posner
2021

TABLE OF CONTENTS

PROFESSIONAL DEVELOPMENT STATEMENT

Since starting the MLIS program in Fall 2019, my ideas of my professional trajectory have shifted, re-focused, and become more realized. I started the program just four months after earning my undergraduate degree in art history, and during the same month that I started MLIS coursework, I also started working as a Library Reference Assistant with the UCLA Arts Library. I began my MLIS journey with the expectation of working as a future art reference librarian or archivist. However, soon after taking the core Description and Access course, I became interested in taking classes focusing on library cataloging, metadata, and technical services. These courses introduced me to another side of being an information professional, and I quickly became motivated to learn the technical aspects of information service. My desire to dive into informatics-based work was still rooted in my core interests—to promote equitable, user-centered information service in the arts and humanities context. However, the primary shift was that I learned how this service could be done from the back end, for example, by designing an accessible website, developing ethical descriptions in catalog records, enhancing a diverse range of digital resources, and so on. The technical proficiencies required for this type of work were very new to me, as I didn't hold any prior advanced technical or computer system skills, so I knew I had some learning challenges ahead of me. Over the next two years, I would become fortunate enough to develop these skills through various learning opportunities, including coursework, internships, and continuing education activities.

The coursework I have completed during this program has had a significant role in my professional development. The two core courses, IS 260 Description and Access and IS 270 Computer Systems and Infrastructures, served as an instrumental

introduction to the practice and theory of library technical services. Furthermore, during my first year, I completed IS 271 Human-Computer Interaction, IS 464 Metadata, IS 461 Descriptive Cataloging, and IS 462 Subject Cataloging and Classification, which were all classes that taught me practical skills that I would later use in future internships. By the end of Spring 2020, I was familiar with library technical services terminology, concepts, and principles and became prepared to apply them.

During the summer between my first and second year, I implemented several coursework lessons in my internship experience as a Junior Fellow with the Library of Congress. For ten weeks, I worked full-time with the Law Library of Congress to help re-design aspects of their public-facing website, work on descriptive cataloging assignments, and develop online research guides for their digital resources. At the same time, I enrolled in an Introduction to Digital Humanities summer course. I learned how to clean datasets, perform text analysis, create data visualizations, and further my skills with web design. I enjoyed the course so much that I would become motivated to earn a Graduate Certificate in Digital Humanities from the UCLA Digital Humanities program. These two experiences over the summer allowed me to practice what I had learned in previous IS courses and become more confident with my technical skills.

I started my second year in the program with a much clearer vision of the kind of informational professional I wanted to be. I knew I could learn what was necessary to become a technical services librarian, and I wanted to continue on that path. To increase my practical experience and align my coursework accordingly, I started a year-long internship with the UCLA Library Cataloging and Metadata Center, continued a Remote Metadata Internship with the Law Library of Congress, and enrolled in the gateway course, IS 214 Informatics. Additionally, I became much more involved in student organizations during my second year. While I was previously involved in

Artifacts and the Special Libraries Association (SLA) student chapter during my first year, I took on board leadership positions as the Web Administrator for Artifacts and as a Co-President for our SLA student chapter. I also wanted to engage in associations beyond the student organizations, so I joined multiple professional organizations, including the Southern California Technical Processes Group (SCTPG) and the Art Libraries Society of North America (ARLIS/NA). I have since contributed to each of these professional organizations. I was elected to be the Secretary and Treasurer for a two-year term with SCTPG, and I will be presenting at the ARLIS/NA annual conference in May and publishing and serving on a subcommittee with ARLIS/NA in the upcoming year. My involvement with these two professional organizations has been advantageous, as I've connected with current specialists in the LIS field and contributed my perspective as a new professional.

Looking ahead, I plan to continue my education in several ways. In the upcoming summer, I will be attending professional conferences and completing educational courses through financial scholarships and awards. I have been selected as the ARLIS/NA 2021 Gerd Muehsam awardee, which will fund my attendance at the annual conference in May, and I am a recipient of a Samuel H. Kress scholarship, which will support my participation in the 2021 Summer Educational Institute for Digital Stewardship of Visual Information in June. I am also applying to the California Rare Book School course, "Digital Humanities for the History of the Book," scheduled for August. As a board member of SCTPG, I will be participating in the upcoming webinars and events. In September 2021, I will begin a position with the Getty Research Institute, where I will work full-time for twelve months. I will be helping with their Digital Art History Initiative to conceptualize and research for the Pre-Hispanic Art Provenance Index. I will be responsible for implementing digital tools to advance the project's

provenance research and data management. With this position, I also have funding available to attend other professional conferences and learning events, which I plan to use during the following year. I am incredibly fortunate to have this range of opportunities ahead of me, which I anticipate will benefit my professional development.

As I near the completion of this program, I reflect on how much I have learned and grown as an early-career professional. My initial expectations have been met with the reality of true development, which required me to get out of my comfort zone, adapt to lessons I didn't foresee and challenge myself to acquire new skills. I am grateful that I have held onto my core interest in promoting arts and humanities research while entering a new technical services context. I am confident that I will continue to grow and lead in this new direction throughout my future endeavors.

Jessica Craig

jessicaecraig11@gmail.com • (805) 670-1590

EDUCATION

Master of Library and Information Science — expected June 2021
Specialization in Informatics
University of California, Los Angeles

Graduate Certificate in Digital Humanities — expected June 2021
University of California, Los Angeles

Bachelor of Arts in Art, Art History emphasis — May 2019
Summa Cum Laude
California State University, Channel Islands

EXPERIENCE

UCLA LIBRARY
Resource Acquisitions and Metadata Services
Cataloging Intern
September 2020 – Present

- Create and enhance original MARC catalog records within OCLC Connexion software client and the UCLA Library's Voyager ILS for incoming foreign-language monographs (Spanish, French, and German).
- Assign descriptive bibliographic metadata to catalog records based on national standards, such as RDA and the Library of Congress Program for Cooperative Cataloging Policy Statements.
- Attach Library of Congress Subject Headings and LC Classification to catalog records for optimal searching and discoverability of collection resources.
- Perform LC-PCC compliant authority work by creating and updating authority files based on NACO and SACO training and guidelines.

UCLA ARTS LIBRARY
Library Reference Assistant
September 2019 – Present

- Provide guidance in the use of physical and digital resources for researchers in the disciplines of fine art, art history, animation, film, television, theater, architecture, museum studies, and design.
- Conduct reference interviews in-person and online to recommend research strategies and sources. Entered and evaluated reference metrics to improve virtual reference services.
- Use digital reference chat platform (LibChat) to assist researchers around the world.
- Created special topic digital research guides (LibGuides) for accessibility on the UCLA Library website.
- Curated a digital exhibit using digital collections and developing a narrative with the use of Adobe Spark.

LAW LIBRARY OF CONGRESS
Digital Resources Division
Remote Metadata Intern
September 2020 – Present

- Lead as the project coordinator; assign metadata projects, review work of 12 interns, assist onboarding and training, ensure organized team workflow and collaboration.
- Parse through digitized U.S. Statutes at Large documents and perform subject analysis for several thousands of enacted laws.

- Assign metadata to each law according to local standards to facilitate their subsequent access on congress.gov.
- Perform metadata clean-up within spreadsheet software for large digital files.

Junior Fellow
May 2020 – July 2020
- Worked to maximize the user experience of the Law Library of Congress website by using graphic and web design principles to promote global access to the library's digital resources and services.
- Enhanced several online research guides for the discovery of digital legal resources.
- Assigned descriptive metadata to digitized historical congressional records to allow for online accessibility.
- Designed collection infographics for the online Legal Reports collection after performing collection data analysis. Presented work to the Law Librarian of Congress.

SANTA BARBARA HISTORICAL MUSEUM
Digital Resource Development Intern
June 2020 – September 2020
- Conducted primary source research to enhance collection provenance records.
- Utilized historical and genealogical digital research tools to gather biographical information about museum collection donors and illuminate women's history in the Santa Barbara-area.
- Recorded and structured completed research into local ArchivesSpace collection management system.

UCLA DEPARTMENT OF INFORMATION STUDIES IS LAB
Research Assistant
September 2019 – March 2020
- Actively performed assistance to student researchers using research lab resources, including the library collections, digital services and software, film and video resource equipment.
- Enhanced collection development based on departmental research areas.
- Managed library collection catalog and processed new acquisitions.
- Coordinated, planned, and led practical research workshops for Library and Information Science graduate students.

ART LIFE FOUNDATION
Archives Catalog Assistant
June 2018 – February 2020
- Collaboratively developed and maintained an original catalog database for the archival collection of rare artist publications and ephemera. Performed metadata quality evaluations regularly.
- Monitored, processed, arranged, described, and researched the physical archival collection.
- Performed best practices for archival preservation of print-based materials, carefully handled and rehoused objects when necessary.

CAMARILLO PUBLIC LIBRARY
Library Assistant II
December 2016 – September 2019
- Assisted a diverse range of community members by sharing helpful information regarding library collection and services based on their individual needs and interests.
- Frequently checked materials for circulation, created library accounts, and solved a variety of account issues.
- Physically handled and processed incoming materials and regularly updated the catalog's bibliographic and item records.

COMPUTER SKILLS & PROFICIENCIES

- **General software and tools:** Microsoft Office (advanced Excel skills), Google Suite, iMovie, Adobe Photoshop, Adobe Illustrator, Adobe Spark, Tableau, OpenRefine, oXygen XML Editor, Confluence, WordPress, Omeka, GitHub
- **Library and archive applications:** Voyager ILS, Polaris ILS, OCLC Connexion, LC Cataloger's Desktop, LC ClassWeb, ArchivesSpace, Springshare LibApps
- **Technical programming:** SQL, XML, HTML, Unix Shell, Git, Python (learning)
- **Metadata management:** Content Standards: RDA, DACS / Structure Standards: MARC, EAD, Dublin Core, VRA Core, MODS / Value Standards: LCSH, LCNAF, LCGFT, AAT, ULAN, TGN

INSTRUCTION / TEACHING

- Teaching Assistant, UCLA Department of Design Media Arts, DESMA 9: Art, Science, and Technology, Spring 2021
- Co-Instructor, "Finding Image Resources Workshop," UCLA Arts Library, April 2021
- Co-Instructor, "Finding Sources in the Library," UCLA Cornerstone Workshop Series, January 2021
- Co-Instructor, "Collecting and Citing Sources," UCLA Cornerstone Workshop Series, October 2020
- Co-Instructor, "Intro to Archival Processing," UCLA Information Studies Research Lab, March 2020

PRESENTATIONS

- ARLIS/NA Annual Conference, New Voices in the Profession session, "Computer Vision for Visual Arts Collections: Looking at Algorithmic Bias, Transparency, and Labor," May 2021 (forthcoming)

PUBLICATIONS

- "Computer Vision for Visual Arts Collections: Looking at Algorithmic Bias, Transparency, and Labor," in *Art Documentation: Journal of the Art Libraries Society of North America* 40, no. 1, Spring 2021

AWARDS

- 2021 Samuel H. Kress Foundation Scholarship
- 2021 Art Libraries Society of North America Gerd Muehsam Award
- 2020 UCLA Information Studies Digital Resource Development Initiative Award

CERTIFICATES

- UCLA Graduate Certificate in Digital Humanities, expected June 2021
- Infopeople Data Privacy Advocacy Certificate, April 2020

VOLUNTEER SERVICE

- J. Paul Getty Museum, Instructional Gallery Docent, October 2018 – April 2021
- Library of Congress, By the People Crowdsourcing Volunteer, January 2020 – May 2020
- iFixit, Cataloging and Metadata Volunteer, April 2020
- CSUCI Alumni Mentorship, Mentor, February 2020 – Present
- Camarillo Ranch House, Tour Docent, June 2016 – October 2018

RELEVANT COURSEWORK

Graduate-level coursework, 2019 – 2021

- **Completed:** Archival Description and Access • Social Science Research Methods • Letterpress Laboratory: Book Arts and Structures • Digital Humanities • Cultural Heritage Preservation • Museums in the Digital Age • Values and Communities in Information Professions • Human-Computer Interaction • Descriptive Cataloging • Subject Cataloging and Classification • Metadata • Resource Description and Access • Media Description and Access • Informatics • Computer Systems and Infrastructures

- **In-progress:** Special Collections Librarianship • Digital Humanities: Designing the User-Centered Art Archive

Undergraduate coursework, 2014 – 2019

- **Completed:** Art History: Tools & Methods • Postmodern Visual Culture • Art, Society, and Mass Media • Multicultural Art Movements • The Museum: Culture, Business • The Business of Art • Visual Technologies • Art of the Ancient World • Medieval Europe 800-1400 • Intercultural Communication • Elementary French

CURRENT PROFESSIONAL INVOLVEMENT

- Art Libraries Society of North America — Student Member
- Artifacts, UCLA Student Organization — Website Admin
- The Horn Press, UCLA Book Arts Student Organization — Member
- Special Libraries Association, UCLA Student Chapter — Co-President
- Southern California Technical Processes Group — Secretary & Treasurer
- Society of American Archivists — Student Member
- Society of California Archivists — Student Member

ISSUE STATEMENT

The effort to utilize machine learning algorithms is becoming widely adopted by visual arts collections as a method to improve access to collections as digitization, born-digital materials, and online open-access efforts continually increase. By spotlighting the high levels of bias and low levels of transparency that often accompany algorithmic object description models, I aim to argue best practices for using machine learning in large visual arts collections.

ISSUE PAPER

Algorithmic Description: Using Machine Learning for Arts-Based Collection Metadata

April 2021

Introduction

Information professionals across various types of collection institutions, libraries, archives, and museums, are often faced with a similar issue. The constant and incredible struggle of an ever-increasing backlog of materials to process, which swells as digitization efforts and born-digital materials proliferate. Limitations of time, funds, and labor continue to block accessibility to collection objects, even as acquisitions may continue. Employment of machine learning for processing digital image collections presents itself as an emerging development and one possible solution to this stagnation. Particularly, computer vision, a subcategory of machine learning, is starting to be implemented in visual arts collections to alleviate long accumulations of unprocessed materials. The use of computer vision in arts-based collections automates digital image analysis and processing to increase metadata description in the hope for its subsequent accessibility. Currently, the growing integration of this new trend in arts collections continues, and broader adoption of the technology appears to be widely embraced by many collection stewards.

From this perspective, the connection of LAM collections and machine learning seems naturally appropriate. The AI technology is apt to process and analyze large amounts of data, like the data held by large collection repositories. Yet, as intrigue for this application of machine learning increases, there are significant issues to address. As recent research studies and literature suggest, machine learning algorithms are known to reflect high levels of bias and low levels of transparency. How can information

professionals working in LAMs confront these potentially harmful effects of machine learning during their use of it? Many institutions that could benefit from automated assistance in collections processing are also institutions which hold equity and inclusivity at the core of their mission. Is there a way to tackle both of these priorities so that institutions can implement responsible collection description and earlier access to materials through the use of machine learning algorithms? Through analysis of recent case studies, interview accounts, and literature, this essay proposes visual arts LAMs can mitigate algorithmic bias by promoting transparency of computer vision models, demonstrating caution, and establishing accountability. The following discussion will introduce computer vision, its implementation in visual arts collections, and follow with the considerations of its ethical application.

What is Computer Vision?

Computer vision is a form of artificial intelligence that automates digital image analysis through a trained machine learning algorithm and convolutional neutral network (CNN) that is able to capture unlabeled images and form judgments about their features and qualities. Computer vision as supervised machine learning means its competency is dependent on the training dataset it learns from, as it makes conclusions based on what it was trained to be correct. The training dataset requires a lot of data, so that the machine learning algorithm can effectively teach itself about the context, similarities, and differences of visual data. The CNN model is a deep learning algorithm that captures the input image, breaks down its pixels, assigns significance based on its learnable weights and biases, makes predictions, and checks the accuracy of those

predictions.[1] As Adam Greenfield explains, "the first goal of machine learning is to teach an algorithm how to generalize. A sound algorithm is one that is able to derive a useful classifier for something it hasn't encountered from the things it has been shown."[2] Once the machine learning algorithm has taught itself to detect patterns, it's able to recognize shapes, objects, colors, as well as more nuanced features, such as faces and sentiment. The application of computer vision for visual art collections in LAMs may help to identify descriptive attributes of a work such as its subjects, style, composition, genre, creator, context, relationships, and even authenticity, with minimum human involvement.[3] If done carefully and conscientiously, the algorithm's ability to identify such characteristics of a digital image can allow it to execute analyses and make certain determinations about it that can source new understandings, reveal connections, and benefit its overall metadata for discoverability and access.[4]

Since the early experimentation of computer vision began in 1959 by two neurophysiologists, David Hubel and Torsten Wiesel, its application has reached far outside of the library and information science field.[5] In 1974, when the technology of optical character recognition (OCR) and intelligent character recognition (ICR) were developed, the possibilities of applying computer vision widened to include documentation and recognition of vehicle plates, mobile payments, invoices, etc.[6] In 2001, the first face recognition applications was introduced by Paul Viola and Michael

[1] Sudeepti Surapaneni, Sana Syed, and Logan Yoonhyuk Lee, "Exploring Themes and Bias in Art using Machine Learning Image Analysis," *2020 Systems and Information Engineering Design Symposium (SIEDS)*, (2020): 1-6, doi: 10.1109/SIEDS49339.2020.9106656.
[2] Adam Greenfield, *Radical Technologies: The Design of Everyday Life*, (London: Verso, 2017), 217.
[3] Brendan Ciecko, "6 Ways That Machine Vision Can Help Museums," *Cuseum* (blog), last modified March 10, 2016, https://cuseum.com/blog/6-ways-that-machine-vision-can-help-museums.
[4] Babak Saleh, Kanako Abe, Ravneet Singh Arora, and Ahmed Elgammal, "Toward Automated Discovery of Artistic Influence," *Multimedia Tools and Applications* 75 (2016): 3565–3591, https://doi.org/10.1007/s11042-014-2193-x.
[5] D. H. Hubel and T.N. Wiesel, "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *The Journal of Physiology* 148 (1959), doi: 10.1113/jphysiol.1959.sp006308.
[6] "Computer Vision," IBM, accessed December 2020, https://www.ibm.com/topics/computer-vision.

Jones,[7] and throughout the 2000s, the procedure of how visual datasets are annotated emerged with the development of new CNN models and image data sets.

Applications

How well different CNNs perform on art and art historical images has been recently researched by numerous scholars, including Sudeepti Surapaneni, Sana Syed, Logan Yoonhyuk Lee (University of Virginia School of Data Science), Sean Yang, et. al. (Univeristy of Washington), and Adrian Lecoutre, Benjamin Negrevergne, Florian Yger. These three studies identify the most extensively used CNNs in image classification tasks as ResNet 50, ResNet 101, Inception-Resnet-V2, and AlexNet, with the latter two displaying the highest performance. Frequently used datasets in these research studies are ImageNet (over 14 million images), The Metropolitan Museum of Art's online collection (375,000 images), WikiArt (140,000 images), and Artsy (27,000 images). In the commercial world, some of the most prominent computer vision tools are being developed and sold by Microsoft, IBM, and Google.

In mid-October 2020, Microsoft announced the recent computer vision developments of their AI product, Azure Cognitive Services, stating, "our Computer Vision image captioning capability now describes pictures as well as humans do."[8] The bold statement was followed by their claims of improved content discoverability, text extraction, image, and video analysis in the effort to advance visual data processing. How exactly Microsoft's product works requires some uncovering and specialized

[7] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2001): 511-518, doi: 10.1109/CVPR.2001.990517.

[8] John Roach, "What's That? Microsoft's Latest Breakthrough, Now in Azure AI, Describes Images as Well as People Do," *The AI Blog, Microsoft*, October 14, 2020, https://blogs.microsoft.com/ai/azure-image-captioning/.

knowledge, as their most public-facing explanation is rather imprecise; vaguely asserting their algorithm "pulls from a rich ontology of more than 10,000 concepts and objects to generate value" without any immediate further details regarding its dataset's source or process.[9]

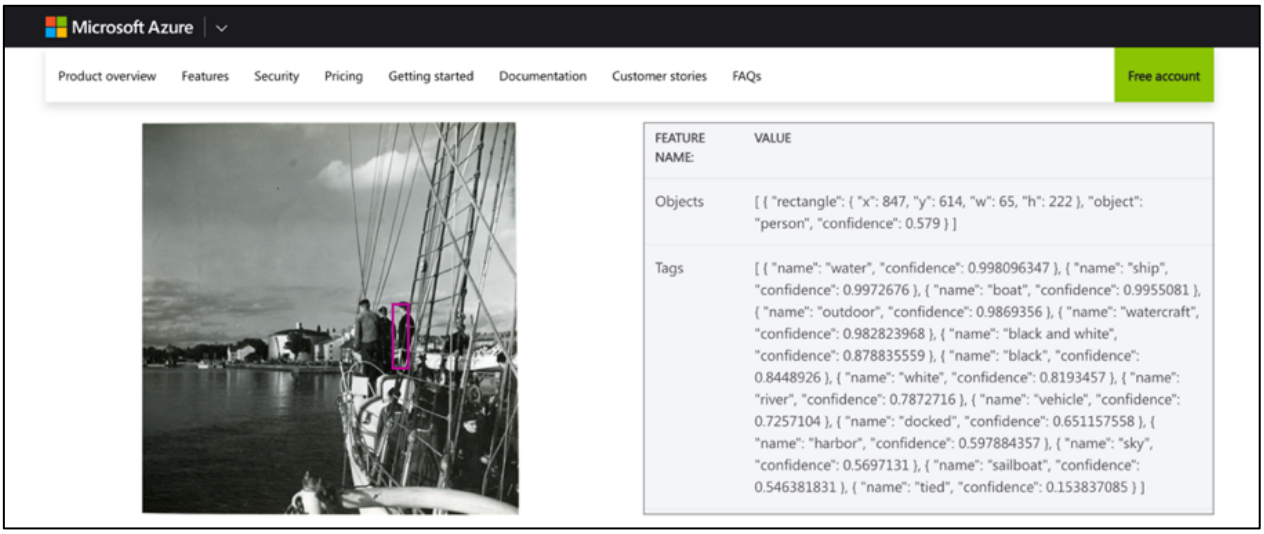


Figure 1. Microsoft Azure results for *Övningsfartyget Gladan anlöper Visby hamn*. 1947. Photograph. Accessed through the DigitaltMuseum. https://digitaltmuseum.se/011014890070.

IBM has also developed computer vision products and services, such as their Watson Visual Recognition product and available code patterns for classifying works of art; although, with much more accessible statements about their methods and with greater emphasis on education. IBM maintains their Trusted AI campaign, "AI Explainability 360", which offers open-source information to guide users in understanding machine learning and elucidate how their models determine and assign labels. "Black box machine learning models that cannot be understood by people are achieving impressive accuracy on various tasks. However, as machine learning is increasingly used to inform high stakes decisions, explainability and interpretability of the models is becoming essential." Although IBM's commitment to transparency and

[9] "Computer Vision," Microsoft, accessed December 2020, https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/#features.

education in AI is a worthy effort, the greatest amount of attention and success has been attributed to the computer vision products developed by a chief competitor in AI innovation, Google.

The concept of computer vision and art has been popularized by the Google Arts and Culture Lab, which has gained remarkable attention for their AI projects using visual collections from approximately 1,000 institutions around the world.[10] Google's computer vision product, Google Vision, has been experimented with by several prominent visual art collections, including by The Met,[11] The Getty,[12] MoMA,[13][14] LACMA[15], and a number of others, such as the Harvard Art Museums, Cleveland

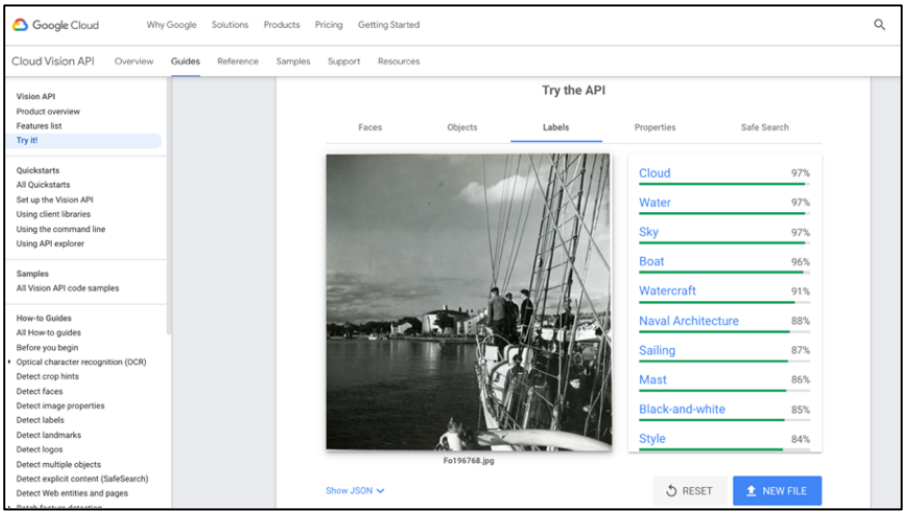


Figure 2. Google's Cloud Vision API results for *Övningsfartyget Gladan anlöper Visby hamn*. 1947. Photograph. Accessed through the DigitaltMuseum. https://digitaltmuseum.se/011014890070.

[10] Amit Sood, "Every Piece of Art You've Ever Wanted to See—Up Close and Searchable," Google Arts & Culture, uploaded June 19, 2016, accessed December 2020, https://www.youtube.com/watch?v=CjB6DQGalU0.
[11] Sarah Robinson, "When Art Meets Big Data: Analyzing 200,000 Items from The Met Collection in BigQuery," *Google Cloud* (blog), last modified August 7, 2017, https://cloud.google.com/blog/products/gcp/when-art-meets-big-data-analyzing-200000-items-from-the-met-collection-in-bigquery.
[12] Nathaniel Deines, "Does It Snow in L.A.? What Computer Vision Saw in Ed Ruscha's Sunset Boulevard," *Getty Iris* (blog), *Getty Museum*, October 7, 2020, http://blogs.getty.edu/iris/does-it-wq2wsnow-in-la/.
[13] "MoMA & Machine Learning," *Experiments with Google* (blog), *Google Arts & Culture*, March 2018, https://experiments.withgoogle.com/moma.
[14] "Identifying Art Through Machine Learning: A Project with Google Arts & Culture Lab" MoMA, accessed December 2020, https://www.moma.org/calendar/exhibitions/history/identifying-art.
[15] Sarah Pham, email message to author, March 26, 2021.

Museum of Art, The Barnes Foundation, and Auckland Art Gallery,[16] establishing it as the most common computer vision tool for visual art collections. For institutions such as these, machine learning—and more specifically, computer vision—has become an aspect of collection management worth supporting and sharing.[17]

Likewise, in academia, we see machine learning and computer vision being heavily researched, worked towards, and applied across a variety of disciplines. In Art History, recent studies by Eva Cetinic, Tomislav Lipic, and Sonja Grgic, "Learning the Principles of Art History with Convolutional Neural Networks" introduces CNN models to predict the visuals features of Heinrich Wölfflin's five key visual principles: linear/painterly, planar/recessional, closed form/open form, multiplicity/unity, absolute clarity/relative clarity.[18] Ahemed Elgammal, et. al. similarly looks at the way machine classification of art styles can relate to art historian's approaches to analyzing style, in their article "The Shape

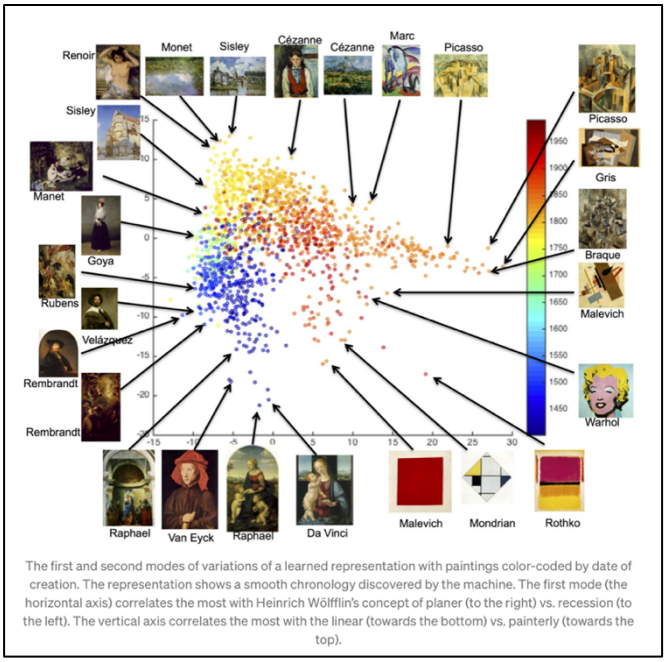


Figure 3. Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone. 2018. "The Shape of Art History in the Eyes of the Machine". *Proceedings of the AAAI Conference on Artificial Intelligence* 32 no. 1.

[16] Brendan Ciecko, "AI Sees What? The Good, the Bad, and the Ugly of Machine Vision for Museum Collections," *The Museum Review* 5, no. 1 (January 2020), http://articles.themuseumreview.org/tmr_vol5no1_ciecko.
[17] Melissa Gill and Nathaniel Deines, interview by Jessica Craig, April 2, 2021.
[18] Eva Cetinic, Tomislav Lipic, and Sonja Grgic, "Learning the Principles of Art History with Convolutional Neural Networks," *Pattern Recognition Letters,* 129 (2020): 56-62. https://doi.org/10.1016/j.patrec.2019.11.008.

of Art History in the Eyes of the Machine."[19] Zhu et. al.'s "Machine: The New Art Connoisseur" and Saleh, et. al., "Toward automated discovery of artistic influence" examine art styles and influence through the lens of machine learning conclusions.[20]

In library and information science (LIS), the topic has emerged to become a dominant area of focus among informatics-based scholars concerned with big data, algorithms, and coded biases, such as Safiya Noble, Thomas Padilla, and Ryan Cordell. Brendan Ciecko describes machine learning have increasingly been established in LAMs in many ways, including analyzing patron use, promoting outreach, and, of course, collection processing.[21] Generally, the ongoing dissemination of case studies has resulted in inspiration within the broader LIS field that stimulates curiosity and interest in computer vision's potential effects for other related collection environments. While its wide implementation is still emerging, current interest is present and rising. A recent OCLC research study conducted by Thomas Padilla from March to August 2019 found that libraries expressed a high level of interest in machine learning and algorithmic solutions for a range of reasons, including increasing efficient collection description, discoverability, and access, along with the prospect of freeing up employee time to meet other shifting demands.[22] While not without concern, however, the idea that a catalog record produced by AI is better than having no record at all was well received.

---

[19] Ahmed Elgammal, Bingchen Liu, Diana Kim, Mohamed Elhoseiny, and Marian Mazzone, "The Shape of Art History in the Eyes of the Machine," *Proceedings of the AAAI Conference on Artificial Intelligence* 32 no. 1 (2018) https://ojs.aaai.org/index.php/AAAI/article/view/11894.
[20] Yucheng Zhu, Yanrong Ji, Yueying Zhang, Linxin Xu, Aven Le Zhou, and Ellick Chan, "Machine: The New Art Connoisseur" *Cornell University arXiv preprint* (2019), https://arxiv.org/pdf/1911.10091.pdf.
[21] Brendan Ciecko, "Examining the Impact of Artificial Intelligence in Museums," *MW17: MW 2017*, last modified February 1, 2017, https://mw17.mwconf.org/paper/exploring-artificial-intelligence-in-museums/.
[22] Thomas Padilla, "Responsible Operations: Data Science, Machine Learning, and AI in Libraries," *OCLC Research*, (December 2019): 6-22, https://doi.org/10.25333/xk7z-9g97.

This prospect of advancing collection accessibility through fast description may seem familiar to those in the archival field, to recall the popular dialog that Mark Greene and Dennis Meissner sparked with their controversial 2005 article in the *American Archivist*, "More Product, Less Process." We've since learned that higher-level description for earlier access is not always in the best interest of the users, workers, or subjects in the collection. But if machine-created description can effectively describe materials at the item-level, it may be a solution that finally solves the long-time MPLP debate among archivists; as the AI-based catalog might reply, "why not both?"[23] However, an answer to such a question requires careful and critical thought about the ranging ethical implications of introducing AI to image-based collections.

While progress and efficiency through computer vision may hold some promise for visual art collection description and access, there are concerns worth addressing regarding its financial expense, impact on labor, and overall effectiveness; although, the two most significant issues discussed here are the probability of algorithmic biases and lack of transparency. Critical inquiry of machine learning and computer vision has found bias and obscurity of algorithmic models to be forefront concerns in immediate need of examination. The missions of LAMs are often centered on principles dedicated to ensuring equitable access and use of their collections. Can machine learning, as a technology known to be biased and opaque, contribute to such an objective? Based on the continual movement towards its adoption in LAMs, there seems to be an affirmative position regarding this, thus a more useful question may be how LAMs can manage these issues to mitigate harm for their collections and users.

[23] Mark Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *The American Archivist* 68, no. 2 (September 1, 2005): 208–263, https://doi.org/10.17723/aarc.68.2.c741823776k65863.

Algorithmic Bias

Computation and technology are often mistaken as neutral and more objective than human-derived thought. However, scholars such as Cathy O'Neil, Joy Buolamwini, Safiya Noble, Kate Crawford, have discredited this misconception with extensive research and examination. Multiple investigative reports have found machine learning algorithms to reflect the beliefs and values of their developers and, consequently, be just as capable of partiality and subjectivity as humans themselves. Whether intended or not, O'Neil describes, "Models are opinions embedded in mathematics."[24] The viewpoint here is not asserting algorithmic bias is problematic just because it can create inaccurate annotations (eg. labelling a static image with house, tree, bicycle, etc.), but rather a deeper, much more severe problem with greater consequences. Biased algorithms are able to perpetuate prejudice, bigotry, and racial profiling, just as humans do. This is clearly explained in Noble's book, *Algorithms of Oppression: How Search Engines Reinforce Racism,* where there are multiple examples of Google's algorithms performing racist search outputs against Black girls and people of color.

As machine learning and computer vision gradually become commonplace in collections, this truth should not be disregarded. The visual collections held by LAMs often hold cultural, social, religious, and political significance, acting as emblems of practice, tradition, and ideology for communities unique in nature and origin. The responsibility of LAMS to care for such objects extends beyond their physical or material state, but also to their digital representation and access. If computer vision as a machine learning technology is inherently biased just as humans are, then the initial

[24] Cathy O'Neil, *Weapons of Math Destruction,* (Broadway Books, 2016).

utilization of it is more likely to be successful if the efforts are spent on managing bias, rather than hopelessly trying to eliminate it. To reduce algorithmic bias for LAMs using computer vision, demonstrating caution and accountability are fundamental first principles of the practice.

The need for caution by LAMs begins at their first point of interaction with computer vision since the operation of it demands one requirement at the start: massive amounts of data to train the algorithm. A crucial note is that datasets are just as prone to bias as the algorithms are, because it is where the bias is derived (after being derived from its creator). If a training dataset is too small or homogenous, what the algorithm can learn will be restricted to the narrow contents of the dataset. The machine learning algorithm only has the ability to perceive what it was taught to learn based on the dataset, so the dataset's contents must be highly varied and diverse. The alternative would lead to a highly biased output. The result of biased datasets can be detrimental, as Ryan Cordell states, "The biases, limitations, and oversights of those datasets will produce flawed research that does not represent the communities libraries seek to serve."[25] To ensure equitable service, the importance of caution should be instilled right away, whether the model is being created or adopted. For computer vision products that allow users to create their own algorithm model (like the ones developed by Microsoft and IBM), attentiveness will be required when choosing and introducing the original training dataset; whereas with products which instead presents their established model for adoption (like Google Vision), the attention will need to be

[25] Ryan Cordell, "Machine Learning + Libraries A Report on the State of the Field," *Library of Congress*, (July 14, 2020): i-86, https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.13.

directed towards investigating the model's pre-existing sources and predetermined make-up.

Examining the authority of established computer vision models could expose insights into what Joy Buolamwini calls, "the coded gaze," which she describes as "a view that posits any technology created by humans will reflect individual or collective values, priorities and if unchecked, prejudices"[26] and situates it as "the embedded views that are propagated by those who have the power to code systems."[27] The coded gaze is a necessary consideration for visual arts collections, which likely hold objects depicting people and faces within their content. The concept of the "looking gaze" is familiar to those in the visual arts, as the feminist theory of the male gaze has dominated art theory, criticism, and movements since Laura Mulvey first coined the phrase in 1975. The coded gaze may also be related to bell hooks' oppositional gaze, which reestablishes the feminist theory as a strategy for Black women to engage critically with mass media representation.[28] Comparably, Buolamwini argues the coded gaze is behind technology applications and models that discriminate against historically marginalized populations; Black women and women of color in particular.[29] There is also evidence of computer vision applications misgendering the subjects of artworks. In results of Surapaneni, Syed, and Lee's research, "the model mislabeled 296 males as females and 363 females as males… a Native American male with long hair, wearing a traditional

[26] Joy Buolamwini, "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers." PhD diss., Massachusetts Institute of Technology, 2017.
[27] Joy Buolamwini "InCoding – In the Beginning Was the Coded Gaze," *MIT Media Lab* (blog) *Medium*, May 16, 2016, https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d.
[28] bell hooks, "The Oppositional Gaze: Black Female Spectators," *The Feminism and Visual Culture Reader* (2003): 94-105.
[29] Safiya Noble, *Algorithms of Oppression,* (New York: New York University Press, 2018).

regalia including a breechcloth which resembles a dress was incorrectly classified as a female given those specific features."[30]

Even while Padilla's research highlights optimistic impacts of machine learning in library collections, he discusses the importance of evaluating positive impacts against the negatives, stating positives "must be weighed relative to a broader field of misuse spanning applications that lack the ability to recognize the faces of people of color, that discriminate based on color, and that foster a capacity for discrimination based on sexuality."[31] Therefore, careful attention to the inter-workings of the computer vision application prior to its adoption is essential to ensure proper use. If adopted, caution is no longer a sufficient sole priority, but should be accompanied by transparency and accountability.

The need for transparency and accountability is another aspect of managing bias in algorithmic models. Unfortunately, as O'Neil's research suggests, transparent models are rare. "Opaque and invisible models are the rule, and clear ones very much the exception."[32] With the lack of transparency, managing bias and accountability remains difficult and problematic. In the collaborative article by Sarah Myers West, Meredith Whittaker, and Kate Crawford, their *Recommendations for Addressing Bias and Discrimination in AI Systems*, states "Remedying bias in AI systems is almost impossible when these systems are opaque. Transparency is essential and begins with tracking and publicizing where AI systems are used, and for what purpose."[33] For LAM institutions

[30] Sudeepti Surapaneni, Sana Syed, and Logan Yoonhyuk Lee, "Exploring Themes and Bias in Art using Machine Learning Image Analysis," *2020 Systems and Information Engineering Design Symposium (SIEDS),* (2020): 1-6, doi: 10.1109/SIEDS49339.2020.9106656.
[31] Padilla, "Responsible Operations," 9.
[32] O'Neil, *Weapons of Math Destruction.*
[33] Sarah Myers West, Meredith Whittaker, and Kate Crawford, "Discriminating Systems: Gender, Race and Power in AI," *AI Now Institute* (April 2019): 4, https://ainowinstitute.org/discriminatingsystems.html.

whose services are often based on inclusive use and participation, there is the potential for new and more transparent methods of machine learning.

Recommendations

When using computer vision for collections description in a LAM context, there are several of ways transparency and accountability can be practically demonstrated. First, inform users that computer vision is used for generating collection metadata and description. As a basic level of transparency, users should be cognizant about the use of computer vision and machine learning for object description. Additionally, providing further information about how computer vision operates in the collection is beneficial for their potential inquiry into the process. Second, information regarding the origin of the training dataset used for the collection should be made easily accessible. Providing readily accessible information about how the computer vision algorithm was trained based on a dataset adds another layer of transparency. Offering educational materials, contact information, and further resources on the explainability of AI in the collection will support their informed use. Third, allow users to control the parameters of the computer vision algorithm applied to their search when exploring a collection.[34] The ability for users to set limitations on the function of AI in their search should be offered. Assumptions about the advantages of AI-generated description should not be forcibly applied to users who may not want to use it in their research discovery process. While this may reduce their searching capabilities, at minimum, it will reveal how much the collection relies on AI for their collection metadata, which allows users the choice to compare and contrast the impact of AI on collection descriptions. Fourth, apply a

[34] Padilla, "Responsible Operations," 10.

tracking feature to object description that allows users to see the alterations of an object's metadata over time. This form of transparency is useful whether the change in the description was caused by a computer or a human, however especially with an algorithmic-induced change, since the algorithmic model should be constantly learning and updating its outputs.[35] This capability contributes to O'Neil's notion of a trustworthy model, "Whatever they learn, they can feed back into the model, refining it… They maintain a constant back-and-forth with whatever world they're trying to understand or predict."[36] This is especially true with the beginning stages of computer vision description in collections since it will likely improve with time. Tracking and publishing those changes keeps the model transparent and its implementation accountable. And fifth, catalog records should list the percentage of certainty assigned to auto-generated metadata descriptions. Revealing the level of certainty given to an auto-tag will benefit how users interpret the description. "Attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measures of certainty."[37] Uncertainty in description is inevitable and knowledge of it is crucial for transparency. The ability for error is present whether the description is assigned by a human or an algorithm, however, the reality of it becomes more apparent when the varying levels of accuracy and certainty are acknowledged.

The practical formation of any or all of these five points will contribute to the transparency and accountability of machine learning and computer vision in LAM visual collections. Each point recognizes the possibility of algorithmic bias but

[35] Padilla, "Responsible Operations," 10.
[36] O'Neil, *Weapons of Math Destruction.*
[37] Padilla, "Responsible Operations," 13.

addresses it by providing ways for users to diminish its impact. Further investigation is required regarding the true effectiveness of these guidelines; however, the result will likely be impactful and set a precedent regarding how LAM collections implement AI for both their collections and user's advantage.

Conclusion

The attention surrounding machine learning and computer vision has burgeoned in the last decade, encouraging several visual arts LAMs to adopt it for their collection analysis and description. Their success with the emerging technology maintains a great amount of potential for broader applications that could benefit efficiency, progress, and workloads in and across LAM institutions. However, the issues of algorithmic bias and transparency should be taken into account by LAMs, both prior to its adoption and afterward. Demonstration of caution and accountability should be evident throughout the process to minimize harmful effects on the collections, workers, and users. With the practical application of such principles and priorities, machine learning and computer vision hold great promise for visual collections in libraries, archives, and museums; but the red flags need not be overlooked to ensure this.

Bibliography

Buolamwini, Joy. "Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers." PhD diss., Massachusetts Institute of Technology, 2017.

Buolamwini, Joy. "InCoding – In the Beginning Was the Coded Gaze." *MIT Media Lab* (blog) *Medium*, May 16, 2016. https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d.

Cetinic, Eva, Tomislav Lipic, and Sonja Grgic. "Learning the Principles of Art History with Convolutional Neural Networks." *Pattern Recognition Letters,* 129 (2020): 56-62. https://doi.org/10.1016/j.patrec.2019.11.008.

Ciecko, Brendan. "AI Sees What? The Good, the Bad, and the Ugly of Machine Vision for Museum Collections," *The Museum Review* 5, no. 1 (January 2020), http://articles.themuseumreview.org/tmr_vol5no1_ciecko.

Ciecko, Brendan. "Examining the Impact of Artificial Intelligence in Museums." *MW17: MW 2017*. Last modified February 1, 2017. https://mw17.mwconf.org/paper/exploring-artificial-intelligence-in-museums/.

Ciecko, Brendan. "6 Ways That Machine Vision Can Help Museums." *Cuseum* (blog). Last modified March 10, 2016. https://cuseum.com/blog/6-ways-that-machine-vision-can-help-museums.

Cordell, Ryan. "Machine Learning + Libraries A Report on the State of the Field." *Library of Congress,* (July 14, 2020): i-86. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig.

Deines, Nathaniel. "Does It Snow in L.A.? What Computer Vision Saw in Ed Ruscha's Sunset Boulevard." *Getty Iris* (blog), *Getty Museum,* October 7, 2020. http://blogs.getty.edu/iris/does-it-wq2wsnow-in-la/.

Google. "Google Arts & Culture." Accessed December 2020. https://artsandculture.google.com.

Google Arts & Culture. "MoMA & Machine Learning." *Experiments with Google* (blog). March 2018. https://experiments.withgoogle.com/moma.

Greene, Mark, and Dennis Meissner. "More Product, Less Process: Revamping Traditional Archival Processing." *The American Archivist* 68, no. 2 (September 1, 2005): 208–263. https://doi.org/10.17723/aarc.68.2.c741823776k65863.

Greenfield, Adam. *Radical Technologies: The Design of Everyday Life.* London: Verso, 2017.

hooks, bell. "The Oppositional Gaze: Black Female Spectators." *The Feminism and Visual Culture Reader* (2003): 94-105.

Hubel, D. H., T.N. Wiesel. "Receptive Fields of Single Neurones in the Cat's Striate Cortex." *The Journal of Physiology* 148 (1959). doi: 10.1113/jphysiol.1959.sp006308.

IBM. "Computer Vision." Accessed December 2020. https://www.ibm.com/topics/computer-vision.

IBM Research Trusted AI. "AI Explainability 360." Accessed December 2020. http://aix360.mybluemix.net/resources#overview.

Microsoft. "Computer Vision." Accessed December 2020. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/#features.

MoMA. "Identifying Art Through Machine Learning: A Project with Google Arts & Culture Lab." Accessed December 2020. https://www.moma.org/calendar/exhibitions/history/identifying-art.

Moriarty, Adam. "A Crisis of Capacity: How can Museums use Machine Learning, the Gig Economy and the Power of the Crowd to Tackle Our Backlogs." *MW19:MW 2019*. Last modified January 15, 2019. https://mw19.mwconf.org/paper/a-crisis-of-capacity-how-can-museums-use-machine-learning-the-gig-economy-and-the-power-of-the-crowd-to-tackle-our-backlogs/.

Ngo, Ton, and Winnie Tsang. "Classify Art Using TensorFlow." *IBM* (blog). October 6, 2017. https://developer.ibm.com/patterns/classify-art-using-tensorflow-model/.

Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: New York University Press, 2018.

O'Neil, Cathy. *Weapons of math destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books, 2016.

"Övningsfartyget Gladan anlöper Visby hamn." DigitaltMuseum. Accessed December 2020. https://digitaltmuseum.se/011014890070.

Padilla, Thomas. "Responsible Operations: Data Science, Machine Learning, and AI in Libraries." *OCLC Research* (December 2019): 6-22. https://doi.org/10.25333/xk7z-9g97.

Roach, John. "What's That? Microsoft's Latest Breakthrough, Now in Azure AI, Describes Images as Well as People Do." *The AI Blog*, *Microsoft*, October 14, 2020. https://blogs.microsoft.com/ai/azure-image-captioning/.

Robinson, Sarah. "When Art Meets Big Data: Analyzing 200,000 Items from The Met Collection in BigQuery." *Google Cloud* (blog). August 7, 2017.

https://cloud.google.com/blog/products/gcp/when-art-meets-big-data-analyzing-200000-items-from-the-met-collection-in-bigquery.

Saleh, Babak, Kanako Abe, Ravneet Singh Arora, and Ahmed Elgammal. "Toward Automated Discovery of Artistic Influence." *Multimedia Tools and Applications* 75 (April 2016): 3565–3591. https://doi.org/10.1007/s11042-014-2193-x.

SAS. "Machine Learning: What It Is and Why It Matters." Accessed December 2020. https://www.sas.com/en_us/insights/analytics/machine-learning.html.

Sood, Amit. "Every Piece of Art You've Ever Wanted to See--Up Close and Searchable." *Google Arts & Culture.* Uploaded June 29, 2016. Accessed December 2020. Video, 15.00. https://www.youtube.com/watch?v=CjB6DQGalU0.

Srinivasan, Ramesh. "Automation Is Likely to Eliminate Nearly Half Our Jobs in the Next 25 Years. Here's What to Do." *Los Angeles Times*, October 29, 2019. https://www.latimes.com/opinion/story/2019-10-29/opinion-automation-is-likely-to-eliminate-40-of-jobs-in-the-next-25-years-heres-what-we-can-do-a.

Surapaneni, Sudeepti, Sana Syed, and Logan Yoonhyuk Lee. "Exploring Themes and Bias in Art using Machine Learning Image Analysis." *2020 Systems and Information Engineering Design Symposium (*2020). https://doi.org/10.1109/SIEDS49339.2020.9106656.

Viola, P., M. Jones. "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2001): 511-518. https://doi.org/10.1109/CVPR.2001.990517.

West, Sarah Myers, Meredith Whittaker, and Kate Crawford. "Discriminating Systems: Gender, Race and Power in AI." *AI Now Institute* (April 2019). https://ainowinstitute.org/ discriminatingsystems.html.

Zhu, Yucheng, Yanrong Ji, Yueying Zhang, Linxin Xu, Aven Le Zhou, and Ellick Chan. "Machine: The New Art Connoisseur." *Cornell University arXiv preprint* (2019). https://arxiv.org/pdf/1911.10091.pdf.

MAJOR PAPER

Linked Data Ontologies for Art Archives: Definitions, Examples, and Challenges

IS 438B: Archival Description and Access

Professor Jonathan Furner

March 2021

Introduction

Linked open data has become an established and yet still emerging trend in the library and information science field. The efforts to adapt collection descriptions for the Semantic Web have impacted metadata professionals' work across libraries, archives, and museum (LAM) collections, even while its wide operation is yet to be fully realized. In each of these domains, there are specific challenges and possibilities of linked open data. In the archival collection context, linked open data could allow for new routeways and exploration of finding aids and archival surrogates. Even more specifically, the adoption of linked open data in the visual art archives context gives rise to new ways of discovering digital artistic resources. However, the nature of the visual collections presents particular challenges that aren't always solved by the generally established linked open data models. As a result, new data models are developed for better, although exclusive, applications for art and rare material metadata. This "solution" may only be creating a larger problem. The proliferation of ontologies and data models may impact the overall interoperability that linked open data is trying to achieve. This paper aims to introduce and define the topic of linked open data, examine past and present projects of linked open data for visual arts archival collections, and look at the future direction of art archive description and access with linked open data as the number of ontologies continues to increase.

As Karen Gracy explains, procedures for archival description are several, diverse, and evolving. "The development of descriptive practice reveals eagerness to explore, assess, and incorporate new technologies to improve documentation, search, retrieval, and use of archival materials."[1] Linked open data is one such development, as it proposes new methods for collection description and access that drastically shift the fundamentals of common metadata procedure. Even as linked open data gains momentum, it was only recently that the established practice of locally creating and storing metadata became disavowed. Allison Mayer frames this outdated practice asserting that "a 'data silo' is a newly-pejorative term for what was once a standard: metadata sets stored locally, in isolation, usually maintained and accessed internally in a given institution."[2] In contrast, the basic principle of linked open data is just what it sounds like: providing better access to structured data by linking it with other related data on the open Semantic Web to decentralize and broaden its overall access. As an introductory note, it's important to acknowledge e a simple but noteworthy distinction; not all linked data is open, and not all open data is linked.[3] To enable linked open data, it needs to be both. It requires meaningful identifiers to be assigned to the named entities in the data that is then openly published to be referenced by others, thereby constructing an open, web-enabled information network. Setting itself apart from other descriptive practices, one of the core objectives of linked open data is to provide a web-based networked capability between related data based on meaningful links in order to promote cross-institutional interoperability and collaborative access to information.

[1] Karen F. Gracy, "Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges," *Archival Science* 15, no. 3 (2015): http://dx.doi.org/10.1007/s10502- 014-9216-2.
[2] Allana Mayer, "Linked Open Data for Artistic and Cultural Resources," *Art Documentation: Journal of the Art Libraries Society of North America* 34 (2015): https://doi.org/10.1086/680561.
[3] Miriam Posner, "What is Linked Open Data?" Miriam Posner, January 7, 2021, video, 18:43, https://www.youtube.com/watch?v=VZBpFiLbi-Y.

What is Linked Open Data?

There are some fundamental definitions required for a basic proficiency of linked open data. As mentioned, the Semantic Web is the general and evolving infrastructure that allows for linked data. It was conceptualized and introduced by Tim Berners-Lee in 2001 as "not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[4] Along with Burners-Lee, it's led by the World Wide Web Consortium (W3C), who are tasked with maintaining the linked open data framework that can be utilized across applications and institutions. The meaningful links that create and build the webbed network are Uniform Resource Identifiers or URIs. Individually, a URI is "a short string that uniquely identifies a resource such as an HTML document, an image, a downloadable file, or a service."[5] Without URIs, linked data would not be possible; they are required to accurately identify and locate data. Burners-Lee emphasizes the role of URIs in his four central rules of linked data:

(1) Use URIs as names for things

(2) Use HTTP URIs so that people can look up those names

(3) When someone looks up a URI, provide useful information, using standards

(4) Include links to other URIs, so that they can discover more things.[6]

Once a URI has been assigned to a data entity, it can be linked through the Resource Description Framework (RDF), which is the standard language that structures data linkages. RDF is the glue of the linked data network or the pathways that allow for the

[4] Tim Burners-Lee, "Linked Data – Design Issues," last modified June 18, 2009, https://www.w3.org/DesignIssues/LinkedData.html.
[5] Murtha Baca, "Glossary," In *Introduction to Metadata*, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata/glossary/.
[6] Tim Burners-Lee, "Linked Data – Design Issues," https://www.w3.org/DesignIssues/LinkedData.html.

relationships between the data. It's designed as triple statements (subject – predicate – object). To express data in triples, an ontology needs to be identified and implemented. In the context of linked data, an ontology is a conceptual model that defines the properties, relationships, functions, and constraints for a specific domain.[7] Some of the most common ontologies for the cultural heritage domain are CIDOC-CRM and the Europeana data model. The issue of the proliferation of ontologies for linked open data will be discussed in later sections. Once the RDF triples are formed and aligned with a specified ontology, the data can be serialized in a number of encoding standards, most commonly RDF/XML, JSON-LD, N-Triples, and Turtle. To retrieve linked data, the RDF-specific query language, SPARQL, is used to query and receive data. One of its distinct characteristics is that "SPARQL focuses on providing 'answers' as opposed to 'documents.' As a result, SPARQL enables deep graph searching across LOD sources and itself returns RDF data, meaning that a SPARQL query is itself a new LOD data source."[8] Defining these mechanical aspects of linked data is important to understanding the basics of how it works. However, as linked data grows and evolves, so are these technical specifications. And increasingly, metadata professionals working with art collections are developing ways to adapt general linked open data specs to their specialized collections. A number of organizations are leading the way to explore such changes, most notably, the American Art Collaborative, the Library of Congress, Europeana, and the Digital Public Library of America. In the following section, the discussion of how linked open data models and ontology specifications have been adapted and reconfigured specifically for visual art collections (CIDOC-CRM à Linked Art; BIBFRAME à ARM Ontology; Europeana à DPLA data model).

[7] Murtha Baca, "Glossary," http://www.getty.edu/publications/intrometadata/glossary/.
[8] Erik T. Mitchell, "Library Linked Data: Research and Adoption," *Library Technology Reports* 49 (2013): 24.

Linked Open Data Projects

As the adoption of linked data capabilities becomes more widespread, there are several dominant institutions that are paving the way. Some initiatives in the LAM field have already completed the transformation of various controlled vocabularies into linked open data, including the Library of Congress Subject Headings (LCSH), Virtual International Authority File (VIAF), DBPedia, and the Getty Research Institute's major vocabulary sets—Art and Architecture Thesaurus (AAT), Thesaurus of Geographic Names (TGN), and Union List of Artist Names (ULAN). These vocabularies are now able to be used as tools for building links across the Semantic Web. They are often used with ontologies for enhanced, more authoritative descriptions for a range of materials.

A remarkable project that demonstrates the possibility of using controlled vocabularies as linked open data is the American Art Collaborative (AAC), which included the Smithsonian American Art Museum (SAAM) and thirteen other institutions. The SAAM/AAC consortium project began in 2014 to form a cooperative environment for building linked open data for American Art, experiment with reconciliation methods, and develop linked open data guidelines for the broader museum field.[9] The three main components of the project display useful ways of practicing linked data methods. First, attempting to map the data to the CIDOC conceptual reference model (CRM) ontology; second, linking the artist data together through the use of linked open data vocabularies (Getty's ULAN vocabulary, DBPedia, etc.); and third, utilizing and exploring the collections through the new linked open data model. Each of these three phases exposed various challenges, which resulted in

[9] Craig Knoblock, et al. "Lessons Learned in Building Linked Data for the American Art Collaborative." In: *d'Amato C. et al. (eds) The Semantic Web –ISWC 2017*. Lecture Notes in Computer Science, 10588. (2017): https://doi.org/10.1007/978-3-319-68204-4_26.

particular suggestions. In particular, mapping the data from thirteen museums to the CIDOC-CRM ontology was a challenge based on the sheer amount of data to work with (which varied in its tidiness and required some data cleaning), but also because the CIDOC-CRM in itself is complicated and requires specialized knowledge. This challenge produced the need for a new target RDF model, Linked Art, which will be discussed further below. The project also found difficulty with creating links between the data with the multiple vocabularies, as the process required machine-automation to assign links followed by an extended period of human attention to verify those links. Lastly, the project needed to test for optimal ways of presenting the linked data for the user interface. By developing their browsing application, they found that exhibiting the primary entities was just as important as exhibiting the relationships among the entities. The CIDOC-CRM model was reconfigured as a simple schema on the front-end interface, and cross references between data were displayed as clear visualizations. Masking the technical complexities of the data was essential, as not to alienate the non-technical user.[10] The recommendations that resulted in this project are formed as clear, practical guidelines. They suggest, "prepare a complete set of data; relate it to an existing or emerging ontology; map it to an open, machine-readable standard, preferably RDF; link it where possible to external hubs of data; and publish."[11] Generally, the AAC project reflects Berners-Lee other set of 5-star criteria for linked open data:

(1) It's available on the web with an open license

(2) It's available as machine-readable structured data

[10] Craig Knoblock, et al. "Lessons Learned in Building Linked Data for the American Art Collaborative." (2017): 264-276.
[11] Allana Mayer, "Linked Open Data for Artistic and Cultural Resources," (2015).

(3) It's in a non-proprietary format (e.g., CSV instead of Excel)

(4) It uses open standards from W3C (RDF and SPARQL) to identify things

(5) All the above, plus: it's linked to other people's data to provide context[12]

Not only is the AAC project an example of how established standards and guidelines can be applied, but one of the most significant aspects of the AAC initiative is the creation of the new data model, Linked Art. Eleanor Fink, the founder and manager of the AAC, explains the need for the target RDF data model was necessary based on disputes regarding how exactly to apply the CIDOC-CRM model to their specific data and the inconsistencies in their data mapping that occurred as a result. Linked Art became their suitable target model, as it presents itself as a compromise of CIDOC-CRM to promote simplified flexibility. "The AAC target model is a profile of the CRM based on the pragmatism that considers its application by multiple museums; accommodations for interoperability with other uses of RDF; can align with other Linked Data projects; and supports the existing online environment… The AAC target model is thus a balance between knowledge representation and ease of use, while it has

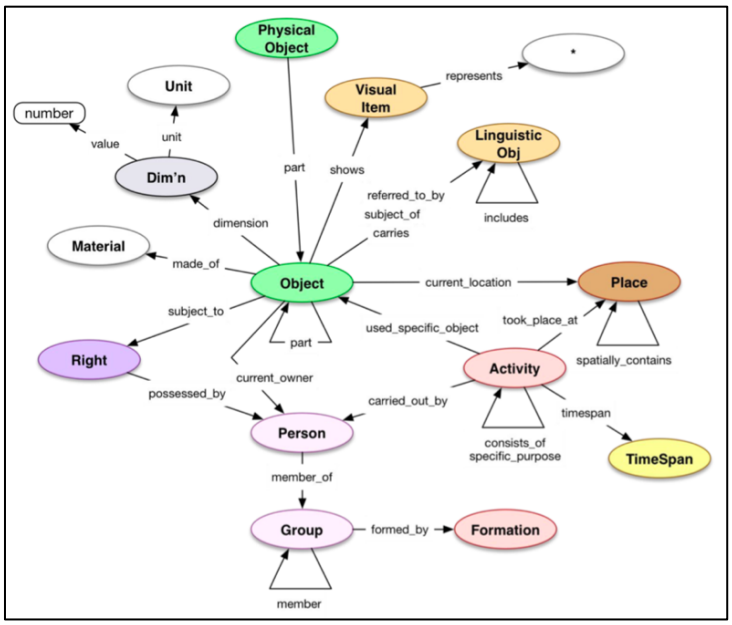


Fig. 1. Linked Art Data Model
Melissa Gill, "LAM Ontologies" class lecture, UCLA GSE&IS, Los Angeles, CA, May 11, 2020.

[12] Tim Burners-Lee, "Linked Data – Design Issues," https://www.w3.org/DesignIssues/LinkedData.html.

the flexibility to accommodate concepts and mappings beyond the target model."[13] Linked Art is founded on the CIDOC-CRM profile; however, it functions with only about 10% of the complexity of the full CRM ontology. The model enables an effective application to be built on top of the model to support varying levels of completeness and be aimed at overall usability.[14] The goal of Linked Art to be applied by multiple museums was not just achieved with the fourteen institutions with the AAC initiative, but it has since gained adoption by several other linked open data developments, such as the Getty Provenance Index, the Linked Conservation Data project (over twenty partnering institutions), the Pre-Raphaelites Art Online project, the PHAROS International Consortium of Photo Archives (fourteen partners)[15], along with several others.[16]

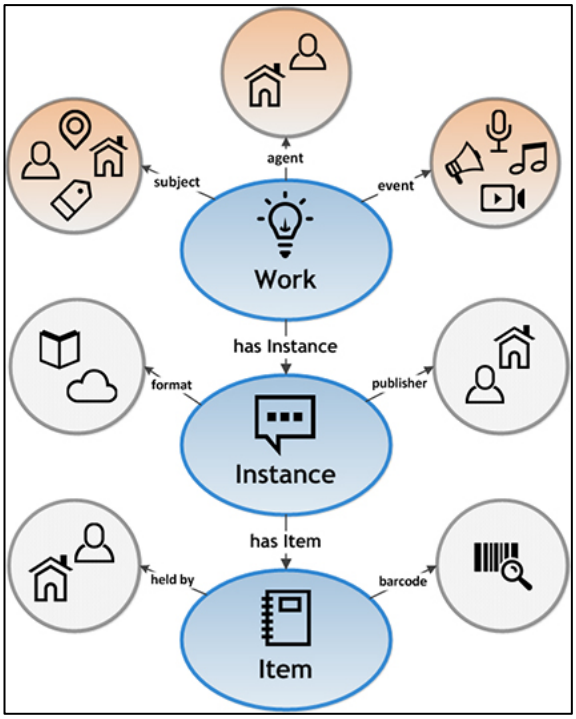


Fig. 2. BIBFRAME 2.0 Model
Library of Congress, April 21, 2016, https://www.loc.gov/bibframe/docs/bibframe2-model.html

Across the board, BIBRAME is one of the most widely anticipated RDF linked data ontologies. It is developed by the Library of Congress and is aimed at the replacement of

[13] Eleanor Fink, "American Art Collaborative (AAC) Linked Open Data (LOD) Initiative: Overview and Recommendations for Good Practices," (2018): 35, https://s3.amazonaws.com/assets.saam.media/files/documents/202007/OverviewandRecommendationsAccessible.pdf.
[14] Melissa Gill, "LAM Ontologies" class lecture, UCLA GSE&IS, Los Angeles, CA, May 11, 2020.
[15] Emmanuelle Delmas-Glass and Robert Sanderson, "Fostering a Community of PHAROS Scholars through the Adoption of Open Standards," *Art Libraries Journal* 45, no. 1 (2020): 19–23. doi:10.1017/alj.2019.32.
[16] "Linked Art Community," Linked Art, accessed March 17, 2021, https://linked.art/community/index.html.

MARC metadata schema to "provide a foundation for the future of bibliographic description, both on the web, and in the broader networked world that is grounded in Linked Data techniques."[17] BIBFRAME is designed to engage with the wider information community, reaching beyond any specific domain in the library, archive, or museum field. The wide application of BIBFRAME is one of the reasons why, similar to CIDOC-CRM with Linked Art, there have been efforts to create a specific extension that's dedicated to for specialized materials. The Linked Data for Production (LD4P) initiative collaboratively develops these BIBFRAME extensions. One of the LD4P projects was Colombia University's ArtFrame project, which focused on developing an extension to BIBFRAME that would be more aligned with the needs of art catalogers. ArtFrame has since been merged with the Rare Materials Ontology to become the Art and Rare Materials BIBFRAME Ontology Extension, or ARM.[18] The ARM Ontology extension allowed the adjustments to BIBFRAME's FRBR conceptual model, which cannot be fully optimized for art and rare materials. In the FRBR model, the "work" is set at the highest-level to represent a disembodied entity, while the physical attributes are described at the lower levels. "This modelization runs counter to museum descriptive practices, in which artworks are regarded as tangible objects."[19] To solve this incongruity, the ARM Ontology created nineteen models. Some models consisting of simple revisions, while others introduce entirely new classes, properties, and relationships. For example, ARM's Marking Model is designed to properly describe an inscribed/printed/stamped/etc. symbol or notation present on a material object. The

[17] "Bibliographic Framework Initiative," Library of Congress, accessed March 17, 2021, https://www.loc.gov/bibframe/.
[18] Elizabeth O'Keefe, Melanie Wacker, and Marie-Chantal L'Ecuyer-Coelho, "The Outcome of the ArtFrame Project: A Domain-Specific BIBFRAME Exploration," *Art Documentation: Journal of the Art Libraries Society of North America* 38, (2019): 9. doi: 10.1086/703508.
[19] Elizabeth O'Keefe, Melanie Wacker, and Marie-Chantal L'Ecuyer-Coelho, "The Outcome of the ArtFrame Project: A Domain-Specific BIBFRAME Exploration," 9.

model consists of eight unique classes (Marking, Autograph, Binder's Ticket, Inscription, Label, Seal, Stamp, and Watermark) and the two properties ('marks' and 'marked by'). In this way, the ARM Ontology addresses the limitations of BIBFRAME in regard to describing item-level characteristics. Similar to MARC, BIBFRAME only contains generic elements for describing physical details (like the 300 $b subfield or 5XX free text fields), so the multiple models of the ARM Ontology alleviate this problem for describing art and rare materials by creating more capability for granular description at the item-level.

> "Several models offer, therefore, a mechanism to identify discrete parts of a resource and describe the distinctive characteristics of each. This allows catalogers to specify, for example, that a book's text block was laser-printed on vellum paper, that its illustrations were painted in watercolors, and that its binding was made of goat suede with leather onlays. Once the information is fragmented and linked to the appropriate resource component, it becomes technically possible to build discovery systems interacting with SPARQL endpoints to enable users to search for objects based on specific criteria—e.g., to look for different examples of bindings made in a certain material or in a certain style."[20]

A key takeaway from ArtFrame project and the creation of the ARM Ontology is understanding the practical ways that linked open data relies on specialized ontologies that are in accordance with the specialized materials they are intended for. Efforts towards segmentation to achieve the appropriate amount of detail should be equated

[20] Elizabeth O'Keefe, Melanie Wacker, and Marie-Chantal L'Ecuyer-Coelho, 13.

with the details of the materials themselves. The conceptualization of the model should be consistent with the conceptualization of the object of description.

One final example of a linked open data ontology for artistic resources is the Europeana Data Model (EDM). Like CIDOC-CRM, the EDM is one of the most relevant and ambitious ontologies for connecting cultural heritage information across libraries, archives, and museums. The EDM draws from a number of existing standards, making it compatible with EAD ad METS standards, aligned with RDA content standard, and several of the EDM descriptive elements are either inspirited by or directly used from Dublin Core, CIDOC-CRM, or SKOS. However, the EDM displays some particular characteristics; for example, it "supports multiple providers describing the same object and allows for enrichment of the museum data, while clearly showing the provenance of all the data that links to digital objects."[21] Also, in contrast to CIDOC-CRM, which is event-driven, the EDM allows for both an object-centric and an event-centric approach. Its main objective is to standardize cross-cultural, multilingual data present. According to a metadata quality report from 2013-2015, Europeana aggregates metadata from over 3,000 cultural heritage institutions and performs enrichment using multilingual vocabularies such as Geonames, Dbpedia and

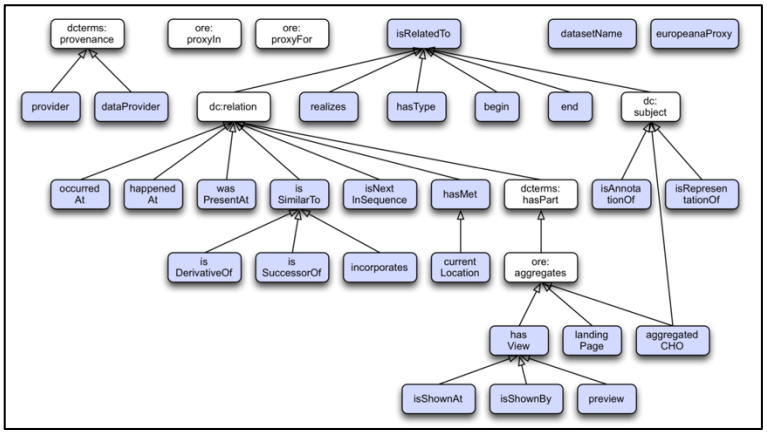


Fig. 3 Europeana Data Model
Europeana, "Definition of the Europeana Data Model v5.2.8, June 10, 2017, https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf

[21] Victoria Boer, et. al., "Amsterdam Museum Linked Open Data," *Semantic Web* 4 (2013): 239. doi:10.3233/SW-2012-0074.

Gemet, as well as other linked open data vocabularies such as the Getty AAT, PartagePlus, Gemeinsame Normdatei (GND), IconClass, and VIAF.[22] One noteworthy adoption of the EDM is by the Amsterdam Museum.[23] The Amsterdam Museum relies on the EDM to maintain their linked open data. The museum ingests and structures their collection metadata for the EDM so that it can be published through Europeana. With cooperation with the EDM, the Amsterdam Museum is able to link to 3,753 external data sources. For scale, Victoria Boer explains, "although this is only a fraction of the total number of concepts, the usefulness of these mappings is much greater as they represent the part of the concepts with which the most metadata are annotated. In total, 70,742 out of the 73,447 (96%) objects are annotated with one or more concepts or persons that have been linked, with an average of 4.3 linked concepts per object."[24] Another effort to adopt the EDM was initiated by the Digital Public Library of America (DPLA). Like Europeana, the DPLA is an aggregator; however, the mission of the DPLA differs slightly from Europeana and certainly from the Amsterdam Museum. The DPLA aggregates metadata, not digital objects themselves, like Europeana does. Thus, the DPLA only uses a fraction of the EDM in order to represent the source content for its discovery. It is focused on linking metadata so that users can be directed to the original external repository outside of DPLA. While the purpose of EDM and DPLA are somewhat aligned as both aggregators across thousands of collections, the DPLA data model that was derived from EDM still stands as another example of a new ontology for selected use.

[22] Marie-Claire Dangerfield, et. al., "Report and Recommendations from the Europeana Task Force on Metadata Quality," *Europeana Think Culture*, (December 2013 - May 2015): 5-20.
[23] "Amsterdam Museum in Europeana Data Model RDF," last modified December 2011, https://semanticweb.cs.vu.nl/lod/am/.
[24] Victoria Boer, et. al., "Amsterdam Museum Linked Open Data," 241.

The challenge of Increasing Ontologies

In 2017, Osma Suomine and Nina Hyvönen, both with the National Library of Finland, wrote the article, "From MARC silos to Linked Data silos?" to discuss the trajectory of linked open data across libraries, archives, and museums.[25] They review the growing trend of libraries to use different data models and argue, "The proliferation of data models limits the reusability of bibliographic data. In effect, libraries have moved from MARC silos to Linked Data silos of incompatible data models."[26] To recall Mayer's critique of data silos, this particular challenge with linked open data is especially problematic.

Through the previously discussed examples of linked data model transformations, from CIDOC-CRM to Linked Art, BIBFRAME's extension to ARM, and Europeana Data Model to DPLA's version, there is evidence to support Suomine and Hyvönen's concern. These are just examples from the arts-based description efforts, but there are more that reach beyond such domain; for example, two of the other most widely used and foundational ontologies from Schema.org and the Simple Knowledge Organization System (SKOS). Further examples are visualized in Suomine and Hyvönen's diagram of data models (fig 4). Similar to the previous discussion, the diagram looks at how the data models stem from each other as a result of prior influence and motivation to create another.

[25] Osma Suominen and Nina Hyvönen, "From MARC Silos to Linked Data Silos?" *O-Bib, Das Offene Bibliotheksjournal / Herausgeber* VDB 4, no. 2 (2017) https://doi.org/10.5282/o-bib/2017H2S1-13.
[26] Osma Suominen and Nina Hyvönen, "From MARC Silos to Linked Data Silos?" 1.

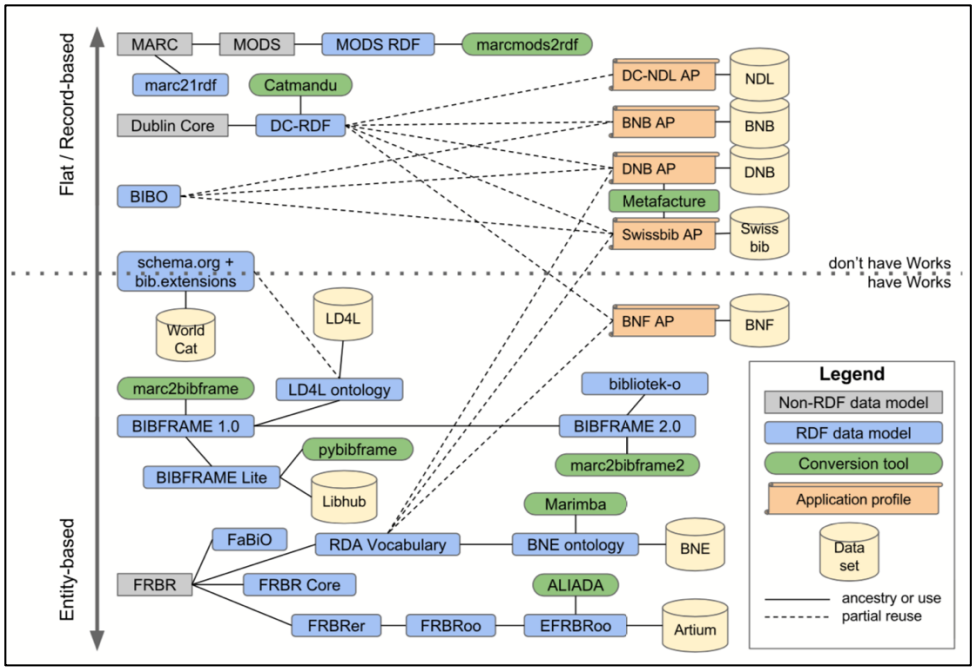


Fig. 4 "Family Forest of Bibliographic Data Models"
Suominen, Osma, and Nina Hyvönen, 2017, "From MARC Silos to Linked Data Silos?" *O-Bib, Das Offene Bibliotheksjournal / Herausgeber* VDB 4 (2):1-13, https://doi.org/10.5282/o-bib/2017H2S1-13.

In order to prevent this issue of too many data models for linked open data, Suomine and Hyvönen propose three solutions. First, try to avoid creating new data models. This recommendation likely sounds easier than it really is, as it requires varied institutions to compromise on common ground. Second, instead of developing new models, effort should be directed towards improving the already established ones. This necessitates institutions to collaborate in a much more active way. Directly working together to make localized and specialized needs known to the greater community could prevent over generalization. "It would help if the collaboration around data models was more open, transparent and organized." One practical way of achieving this is being conscientious of the tools being used to collaborate. GitHub is an example of an open platform that allows for wide input and transparent modes of communication. This was seen with the first project discussed in this paper when the

AAC developed their linked data model that would be subsequently implemented across fourteen institutions; GitHub also used among the communities working with LD4P, RDA Vocabularies, Schema.org, and many more.[27] Finally, Suomine and Hyvönen suggest the possibility of using an externally imposed data model. "a powerful external actor, such as Facebook or one of the major Web search engines, starts harvesting bibliographic data from libraries en masse." This organization would establish the exact representation that libraries would be required to use. "If the service that is based on this harvested data is attractive enough for the libraries, they would have no choice but to provide their bibliographic data using the externally imposed data model, regardless of how difficult this may be for them and how much data quality will suffer in the conversion."[28] If enacted, this final suggestion would replace the cooperative collaborations among institutions and leave little to no room for compromise. Fortunately, this situation is unlikely to occur; it is both interesting and unsettling that the authors would propose it. Giving control over to a larger dominant agent may not be in the best interest LAMs, especially when the authors frame it by saying it would be a result of the absence of choice. This would likely put specialized collections like art archives at an even greater disadvantage as they try to keep up with the more encompassing data models. The first two out of the three suggestions should be more regarded over the third one. Additionally, as demonstrated with the three project examples, the first suggestion of "avoiding new data models" may not be the most feasible. Thus, the second recommendation of improving the established models may be most beneficial for art-based archived aiming to enter the linked open data

[27] Osma Suominen and Nina Hyvönen, "From MARC Silos to Linked Data Silos?" 1.
[28] Osma Suominen and Nina Hyvönen, 12.

world, although a balance between the first and second recommendation is likely more probable.

Conclusion

The efforts towards linked open data are still shifting, adapting, and learning. While there are technical frameworks and conceptualizations in place, there are significant pathways still to be forged, especially in regard to linking artistic resources on the Semantic Web. Through the three examples of Linked Art, ARM Ontology, and DPLA, it is evident that cultural heritage materials require domain-specific provisions in the linked open data initiatives. However, we should be careful to specialize the data models so much that they aren't interoperable for the larger linked open data scheme. A balance of specialization and appropriate generalization may be most beneficial moving forward. However, most importantly, with cooperative and cross-institutional access as the primary goal of linked open data, then collaboration should also remain at the center of the developing process.

Bibliography

"Amsterdam Museum in Europeana Data Model RDF." Last modified December 2011. https://semanticweb.cs.vu.nl/lod/am/.

Baca, Murtha. "Glossary." In *Introduction to Metadata*, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata/glossary/.

"Bibliographic Framework Initiative." Library of Congress. Accessed March 17, 2021. https://www.loc.gov/bibframe/.

Boer, Victoria, et. al. "Amsterdam Museum Linked Open Data." *Semantic Web* 4 (2013): 239. doi:10.3233/SW-2012-0074.

Burners-Lee, Tim. "Linked Data – Design Issues." Last modified June 18, 2009. https://www.w3.org/DesignIssues/LinkedData.html.

Dangerfield, Marie-Claire, et. al. "Report and Recommendations from the Europeana Task Force on Metadata Quality." *Europeana Think Culture* (December 2013 - May 2015).

Delmas-Glass, Emmanuelle, and Robert Sanderson. "Fostering a Community of PHAROS Scholars through the Adoption of Open Standards." *Art Libraries Journal* 45, no. 1 (2020): 19–23. doi:10.1017/alj.2019.32.

Fink, Eleanor. "Overview and Recommendations for Good Practices," *American Art Collaborative (AAC) Linked Open Data (LOD) Initiative* (2018): 35, https://s3.amazonaws.com/assets.saam.media/files/documents/202007/OverviewandRecommendationsAccessible.pdf.

Gill, Melissa. "LAM Ontologies." class lecture. UCLA GSE&IS. Los Angeles, CA. May 11, 2020.

Gracy, Karen F. "Archival description and linked data: A preliminary study of opportunities and implementation challenges." *Archival Science* 15, no. 3 (2015): 239–294. http://dx.doi.org/10.1007/s10502-014-9216-2.

Knoblock, Craig, et al. "Lessons Learned in Building Linked Data for the American Art Collaborative." In: *d'Amato C. et al. (eds) The Semantic Web –ISWC 2017*. Lecture Notes in Computer Science, 10588. (2017): https://doi.org/10.1007/978-3-319-68204-4_26.

"Linked Art Community." Linked Art. Accessed March 17, 2021. https://linked.art/community/index.html.

Mayer, Allana. "Linked Open Data for Artistic and Cultural Resources." *Art Documentation: Journal of the Art Libraries Society of North America* 34 (2015): https://doi.org/10.1086/680561.

Mitchell, Erik T. "Library Linked Data: Research and Adoption." *Library Technology Reports 49* (2013).

O'Keefe, Elizabeth, Melanie Wacker, and Marie-Chantal L'Ecuyer-Coelho, "The Outcome of the ArtFrame Project: A Domain-Specific BIBFRAME Exploration," *Art Documentation: Journal of the Art Libraries Society of North America* 38, (2019). https://doi.org/10.1086/703508.

Posner, Miriam. "What is Linked Open Data?" Miriam Posner. January 7, 2021. Video, 18:43. https://www.youtube.com/watch?v=VZBpFiLbi-Y.

Suominen, Osma, and Nina Hyvönen. "From MARC Silos to Linked Data Silos?" *O-Bib, Das Offene Bibliotheksjournal / Herausgeber* VDB 4, no. 2 (2017). https://doi.org/10.5282/o-bib/2017H2S1-13.

CORE COURSE PAPER

Evaluating Metadata Standards for Cultural Heritage Materials

IS 260: Description and Access

Professor Gregory Leazer

December 2019

Library and information professions are constantly wrapped up in metadata – how it is structured, how it will be used, who creates it and what it really means. Metadata defines much of the world around us and is, therefore, a continual topic of examination. The significance of metadata for an information professional has a lot to do with where it concludes and who it will affect. Our current metadata standards impact various communities in different ways depending on the effectiveness of its structure. Metadata standards, such as the Dublin Core, have proven to contain biases that undermine based on cultural and linguistic systems. The following discussion will critically examine how metadata operates, how it functions to regard cultural data differently, and more specifically how Dublin Core's structure has failed to serve Indigenous communities.

Defining metadata may begin at the most basic level as "data about data." However, metadata and its functions are far more complex. Just in dissecting the simple phrase, we must further define what is meant by "data" and what is meant by "about." Data is the ubiquitous "stuff" all around us; as Jeffrey Pomerantz explains, it is unprocessed *potential* information.[1] However, as information scientists continue to define data, we can be sure that data is a complicated and tenuous concept. In regard to exploring what metadata is, the "aboutness" becomes arguably more relevant and

[1] Jeffrey Pomerantz, *Metadata*. (Cambridge: MIT Press, 2015), 21.

important. The practice of describing an object's aboutness, as simple as it initially sounds, is what cataloguers and metadata specialists must consistently grapple with. The description that is produced becomes metadata, which Pomerantz concludes is "a statement about a potentially informative object."[2]

There are different types of metadata, and within those types, varying standards and uses. Metadata in general functions to discover resources, control intellectual property rights, and among other purposes, to manage, certify the authenticity, identify versions, indicate status, and mark the content structure of documents.[3] Typical types of metadata are descriptive, administrative, structural, and other types serve for preservation, meta-metadata, and more. Out of these, descriptive metadata is the most standardized and used type, as it is used in the traditional library catalog and is accepted as the optimal one for the role of resource discovery. The value standards and controlled vocabularies of descriptive metadata include the Library of Congress Subject Headings and the Library of Congress Name Authority File, among others. These controlled vocabularies are utilized for data structures and element formats, which for descriptive metadata include standards such as MARC 21 (an international standard syntax), Metadata Object Description Schema and Metadata Authority Description Schema (XML standards that uses language-based tags), and the data structure which will be the focus of the following discussion, Dublin Core. Other types of metadata, such as administrative or structural metadata, will provide information to help manage a document and its preservation by describing its technical characteristics, access rights and restrictions, its digital provenance, authenticity, preservation activity, and

[2] Pomerantz, 26.
[3] Hillman, D., Guenther, R., and Hayes, A., "Metadata Standards & Applications Trainee Manual," Library of Congress Workshop Course Materials, last modified August 2008, https://www.loc.gov/catworkshop/courses/metadatastandards/pdf/MSTraineeManual.pdf, 18.

environmental requirements.[4] It is necessary to consider the type of metadata because it will generate different modes of description which will affect how it is retrieved in the future.

An exploration of metadata standards, how they compare and what their flaws are, increases literacy and understanding of description data and their impact on access. Metadata standards can be defined as "the set of fields, words, elements and/or principles for describing resources that are considered to be common to all resources of a particular type – they are inherently universalist and homogenic."[5] Standards are intended to organize the vast and infinite bank of human knowledge, and therefore assume a high level of universality. This is a key discrepancy in the reality of knowledge organization; for instance, such as the case when we consider standards for classifying Indigenous and non-Western sources. Standards are widely accepted rules for producing data about objects, and as these rules create units of information, they work to permeate across distances and different modes of description.[6] "In defining 'standard,' most scholars within the information studies field emphasize the importance of values standards promote, such as compatibility and interoperability, along with what makes information and metadata shareable, searchable, filterable, and retrievable."[7] Due to the effort for interoperability, the means to access information across varying computerized systems for exchange, the standards are created in a manner that cannot be altered easily. Montenegro proposes the question that frames this issue, "how tensions between a western desire for more universal access through

[4] Hillman, D., Guenther, R., Hayes, A., 105.
[5] María Montenegro, "Subverting the Universality of Metadata Standards." *Journal of Documentation* 75, no. 4 (2019): 735.
[6] Geoffrey Bowker and Susan Star, *Sorting Things Out: Classification and Its Consequences* (Cambridge: MIT Press, 1999).
[7] Montenegro, 735.

interoperability can be balanced against the needs of Indigenous communities for localized and culturally responsive documentation and description tools?" Bowker and Star further point out the problematic fact that every standard which becomes successful is wrapped up in the values of the institution that created it. The theoretical neutrality of standards, in fact, develops an intersection of social organization, moral directive, and mechanical cement. Melissa Adler calls this concept, "fixed subjectification," as standards become instruments of dominance that establish what cultural names, titles, categories are validated.[8] Standards are used as a device to reinforce the idea that the creator (or, more likely in some cases, creators) of the information being documented cannot describe their data through the lens of their specific values and beliefs. This creates a disconnect from the metadata and the raw data itself. The following will discuss how the Dublin Core as a metadata structure introduces and perpetuates the "fixed subjectification" of Indigenous knowledge through the assumptions underlying its operation that are connected with cultural and linguistic systems.

The Dublin Core is one of the most widely used metadata schemas. It involves fifteen main elements that are designed to supposedly be able to describe any digital resource.[9] The key feature of the Dublin Core is the way it deliberately functions on the lowest common denominator level. The primary objective of the fifteen-element set is to generate terms that are wide and vague enough to optimize accessing, searching, locating, and retrieving a range of resources that span across knowledge organization systems. Since the establishment of these core elements in 1995, there have been

[8] Melissa Adler, *Cruising the Library: Perversities in the Organization of Knowledge* (New York: Fordham University Press, 2017).
[9] The fifteen elements are contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, and type.

additions to extend its mechanism by the use of a set of terms (for instance, *modified*, *hasPart*, *isPartOf*, *audience*, and many others), the inclusion of qualifiers that allow for a narrower refinement of the element, and certain encoding schemes to facilitate the interpretation of an element value.[10] Nonetheless, it is known for its simplification and flexibility for the purpose of discovering. Specific controlled vocabularies are recommended when entering elements, although not required. The intent was to create simple, low-cost, and easy to use schema in order for its wide acceptance and use. Due to the theoretical ability to describe anything on the digital platform, the Dublin Core shapes various types of information. Not everything being described will offer a value for each of the fifteen elements, in fact, it is mostly the case that some elements will be left blank in the record. The object being described will guide how well the Dublin Core performs, based on the level that it optimizes the fifteen elements, though that is not to minimize the accountability placed on the structure. A major downfall of the Dublin Core is its generalization of data. It does not allow for an adequate level of specificity. The idea of specificity in the field of information relates to Manulani Aluli Meyer's claim that "specificity leads to universality."[11] Marisa Duarte and Miranda Belarde-Lewis expand off this concept in their work "Imagining: Creating Spaces for Indigenous Ontologies" by explaining how knowledge organization work has been heavily impacted by colonialism by historically prioritizing the opposite, that is, generalization and simplification lead to universality. Their proposed theory of imagining is an effort for building Indigenous knowledge systems with the pursuit of decolonization as a

[10] Pomerantz, 65-84.
[11] Manulani Aluli Meyer, "Indigenous and Authentic: Hawaiian Epistemology and the Triangula- tion of Meaning," in *Handbook of Critical and Indigenous Methodologies*, ed. Norman K. Denzin, Yvonna S. Lincoln, Linda Tuhiwai Smith (Los Angeles: Sage, 2008), 217–232, quoted in Marisa Duarte and Miranda Lewis, "Imagining: Creating Spaces for Indigenous Ontologies" *Cataloging & Classification Quarterly* 53, no. 5 (2015): 678.

driving factor. Imagining is a pertinent theory in the discussion of Dublin Core because of the way the data structure works to undermine Indigenous knowledge and discounts its proper organization.

Within some of the fifteen elements, the Dublin Core certainly holds assumptions and biases that relate to cultural systems. This is demonstrated by Dublin Core's definition of the elements and how they are intended for use. In the research of María Montenegro, *Subverting the Universality of Metadata Standards*, it is emphasized how problematic Dublin Core's Creator field and the Rights field are with organizing Indigenous knowledge. Under the Rights field, which is defined as "encompassing Intellectual Property Rights, Copyright and various Property Rights," there is the extended value of the RightsHolder, which is explained as "a person or organization owning or managing rights over the resource." This clearly emphasizes the western concept of ownership and singularity. The Creator field is defined as "an entity primarily responsible for making the content of the resource." These definitions maintain colonial ways of organizing and practices of exclusion by ignoring the various possible systems of ownership and creation. The fields, which are supposedly adaptable to literally any resource, disagree with Indigenous values and belief systems, as Montenegro explains, "the definition provided by Dublin Core for the rights element presumes that IP laws are universal, however, legal regimes of IP and copyright are culturally specific and the types of rights they specify, by definition, exclude all types of Indigenous traditional knowledge."[12] The assumptions regarding IP laws are particular to newly created so-called "original" material created by individual authors. This diverges from Indigenous values as their cultural information is not always new, but

[12] Montenegro, 737.

rather draws upon previous knowledge from ancestral and cultural traditions. Moreover, the concept of a distinct creator (one person, one organization, one service) assumed by the Dublin Core is not aligned with all practices of creating by Indigenous communities, since the work is not necessarily attributed to a single entity but involves collective credit. "One of the fundamental differences between dominant Western and Indigenous knowledge is that the dominant paradigm is based upon the central belief that knowledge is a discrete entity that can be gained and owned by an individual."[13]Additionally, western IP laws and the Dublin Core's creator field maintain the assertion that the creator of the information is the person who was responsible for its documentation.

> "For instance, a film of a traditional ceremony recorded by an ethnographer makes the filmmaker the "author," while the subjects of these colonial documentation practices are rarely given that status. As the "subjects" of these materials instead of the legal copyright owners, Indigenous communities have often no control over the life of their belongings, including in which repository they end up and how they are documented, shared, accessed and used. Furthermore, ironically, they must secure permission from the "author" in order to reuse the materials that document their own lives, customs, and cultural practices."[14]

Therefore, by the lack of attention to cultural specificity and regard for true collective authorship and creation, Indigenous communities are often found without the rights

[13] Lala Hajibayova and Wayne Buente, "Representation of Indigenous Cultures: Considering the Hawaiian Hula", *Journal of Documentation* 73, no 6 (2017): 1139.
[14] Montenegro, 738.

over their production of knowledge. The Dublin Core is only one of the various pathways of this injustice, however, the elements and definitions of those elements they uphold are responsible. While the problematic fields relating to property and ownership disenfranchise Indigenous communities, the ability to describe their information in their linguistic terms further marginalizes them and disassociates the metadata record with the truth.

How Dublin Core, as a classification and data structuring system, operates also relates to issues of linguistic systems. In description practices, language shapes how knowledge is represented and thus is organized. Representation is, therefore, a product of language, D.C. Blair claims "the process of representing documents for retrieval is fundamentally a linguistic process."[15] The relationship between the chosen word and the object, event, or action is a subjective associative bond, as linguist Ferdinand Saussure simply states when discussing his principle of the Arbitrary Nature of the Sign, "the bond between the signifier and the signified is arbitrary."[16] In creating the metadata for Indigenous knowledge work, this becomes a key concern, as Hope Olson explains the significance of representation is a process of naming. "Naming is the act of bestowing a name, of labelling, of creating an identity. It is a means of structuring reality. It imposes a pattern on the world that is meaningful to the namer. Each of us names reality according to our own vision of the world built on past meanings in our own experience."[17] In the case of classifying Indigenous sources of knowledge, typically

[15] D.C. Blair, *Language and Representation in Information Retrieval*. New York: Elsevier Science Publishers, 1990. quoted in Hajibayova, L. and Wayne Buente, 1140.
[16] Ferdinand Saussure, *Course in General Linguistics*, trans. Wade Baskin (New York: Philosophical Library, 1959), 67.
[17] Hope Olson, *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Dordrecht: Kluwer Academic Publishers, 2002. quoted in Hajibayova, L. and Wayne Buente, 1140.

they are managed using national languages, such as English, and thereby molded by the major language, as opposed to the localized language it was originally produced in. This has a major consequence, when Indigenous information is translated and managed "according to western and universalist documentation and classification systems, ignoring and disavowing Indigenous ontologies, epistemologies and local language ideologies."[18] Not only is the language terms lost in translation, but the culturally specific values and beliefs which were extended from it.

The Dublin Core holds underlying assumptions of data based on Western constructions of organization and access. This is furthermore displayed by the static nature of the classification system. Although this may not be particular to Dublin Core, but also applied to other metadata standards, the simple practice of inserting cultural knowledge into a data structure freezes the information from a single time. There is no possibility to place the data into an adaptable or temporal space, and in the case of Dublin Core, it requires inputting by reducing the information into fifteen values to define it. These metadata values, therefore, sentence the data to an unchangeable and therefore inconsistent record. "These practices are often only conducted once by a museum, archives, and library specialists, disregarding the fact that Indigenous knowledge — like any other knowledge — is dynamic and in a constant state of change, depending on the social and cultural flexibility and sustainability of each Indigenous community."[19] The possibility of a living archive may be better suited to account for knowledge which continues to be altered and updated, however any likelihood for a value-free standard continues to be contested.

[18] Montenegro, 734.
[19] Montenegro, 734.

The possibility of a neutral metadata standard remains unlikely. Our present systems, like Dublin Core, among others, have become well adopted largely due to their simplification methods. And in the case where a standardized system does account and include localized forms of meaning within its data, the risk of losing interoperability is a major concern by institutions. Yet, Montenegro proposes another possibility, perhaps the harsh and unfortunate reality that it is based on the apprehension of some professionals within the information field. "Making information standards more flexible has more to do with a profound fear around making space for the voices of other, less privileged and marginalized communities that might challenge the authoritativeness of their discourses around information documentation, and undermine their power and authority to identify, describe and interpret others' materials."[20] The likelihood of this reason for the lack of an adequate metadata standard may be argued for or against, the possibility of one may remain outside of a choice. "Each standard and each category valorizes some point of view and silences another. This is not inherently a bad thing—indeed it is inescapable. But it is an ethical choice, as such, it is dangerous—not bad, but dangerous."[21] As Bowker and Star explain it, standards are fundamentally hierarchal, but it is a fact that can be dealt with by making precautious and ethical decisions.

Metadata may be simplified as "data about data," however, as this paper suggests, simplification does not necessarily mean an understanding of. The theory and practice of metadata are some of the key topics within the information field and will continue to be regarded as complex frameworks for how we access the world's

[20] Montenegro, 738.
[21] Bowker and Star, 5.

knowledge. However, as its practicality has proven, it has a profound impact on how communities can access certain information, even their own. As the possibility for value-free metadata continues to be discussed within the field, the Dublin Core remains as a key example of how metadata elements can define what information becomes valued on a wider scale.

Bibliography

Adler, Melissa. *Cruising the Library: Perversities in the Organization of Knowledge.* New York: Fordham University Press, 2017.

Bowker, G. and S. Star. *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press, 1999.

Duarte, M. and Miranda Lewis. "Imagining: Creating Spaces for Indigenous Ontologies" *Cataloging & Classification Quarterly* 53, no. 6 (2015): 677-702.

Hajibayova, L. and Wayne Buente. "Representation of Indigenous Cultures: Considering the Hawaiian Hula." *Journal of Documentation* 73, no. 6 (2017), 1137-1148.

Hillman, D., Guenther, R., and Hayes, A., "Metadata Standards & Applications Trainee Manual," Library of Congress Workshop Course Materials, last modified August 2008, https://www.loc.gov/catworkshop/courses/metadatastandards/pdf/MS TraineeManual.pdf.

Montenegro, María. "Subverting the Universality of Metadata Standards." Journal of Documentation 75, no. 4 (2019): 731-749.

Pomerantz, Jeffrey. *Metadata*. Cambridge: MIT Press, 2015.

Saussure, Ferdinand. *Course in General Linguistics*. Translated by Wade Baskin. New York: Philosophical Library, 1959.

ELECTIVE COURSE PAPER

Using IFLA's General Principles to Evaluate Moving Image Metadata Schemas

IS 289: Media Description and Access

May Haduong

December 2020

The process of cataloging moving images can take many different forms, largely depending on the metadata standards being used. Standards for the data's structure, value, content, and encoded data format can vary widely, and the differences between them can result in contrasting levels of effective description and access. An analytical look into the standards can reveal the advantages and disadvantages for the collection, its catalogers, and most importantly, its users. This paper will look at metadata structure standards in particular, through a critical analysis of common metadata schemas for cataloging moving images—MARC, Dublin Core, PBCore, and EN 15907—using the IFLA's general principles as criteria to expose the varying levels of the standard's quality and effectiveness for description and access.

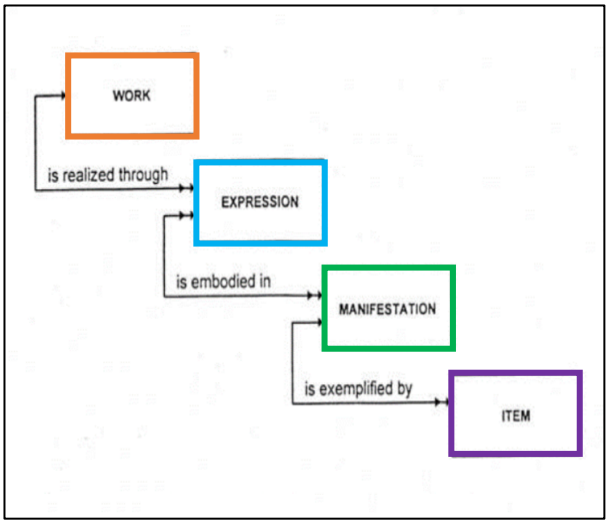
Data structure standards, also described as metadata schemas and element sets, control the organization of metadata by functioning as "containers" of the metadata information.[1] The choice to implement one schema over another usually depends on the materials being described and the nature of intended access. For this paper, the decision to look into these four metadata schemas is based on the different features of each standard and the desire for variation in their origins and purposes; two schemas are designed for wide application (MARC and Dublin Core), whereas the other two are

[1] Anne Gilliland, "Setting the Stage." In *Introduction to Metadata*, edited by Murtha Baca, 3rd ed. Los Angeles: Getty Publications, 2016, http://www.getty.edu/publications/intrometadata/setting-the-stage/.

intended for audiovisual materials specifically (PBCore and EN 15907). The structure of each schema is distinct, based on the intended purposes and concepts that formed its creation.

MARC, known as the machine-readable catalog, is one of the most widely seen metadata standards in and across collection catalogs. The use of MARC can be complex, although its purpose is simple. It is intended to allow for bibliographic, authority, and holdings records to be read by a machine and consequently become intelligible to users. In many ways, MARC supports the Functional Requirements for Bibliographic Records (FRBR) conceptual data model (Figure 1), meaning that it allows for a work, expression, manifestation, and item to be expressed in a single record (see Table 1). Two key aspects of MARC most relevant to this analysis are, first, that MARC is not designed for a specific format but instead can be used broadly regardless of the information object's carrier; and second, it is primarily aimed for machine comprehension and then for human comprehension.

Figure 1. **FRBR Group 1 Entities and Relationships**



IFLA Study Group on the Functional Requirements of Bibliographic Records, "Functional Requirements of Bibliographic Records: Final Report," *International Federation of Library Associations and Institutions*, September 1997, http://www.ifla.org/VII/s13/frbr/frbr.pdf.

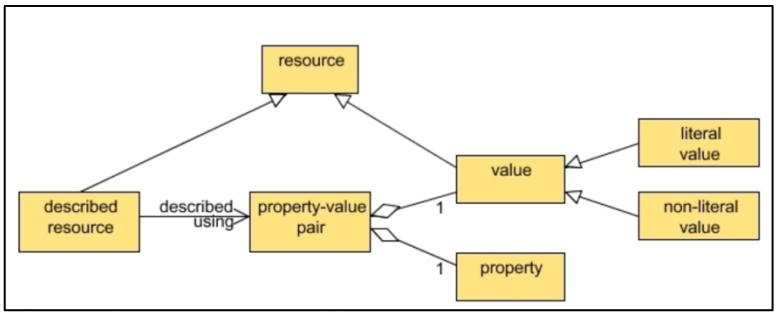
Table 1. **FRBR Group 1 Entities Expressed in MARC Records**

MARC Field	FRBR Group 1 Entity
1XX	work, expression
240	work, expression
245 - 260	manifestation
300	expression, manifestation
3XX	work, expression, manifestation
490	manifestation
5XX	work, expression, manifestation, item
700 - 730	work, expression
760 - 787	work, expression, manifestation
8XX	work, expression, manifestation

Table devised using "FRBR: FRBR, RDA, and MARC," *Library of Congress*, September 2012, https://www.loc.gov/catworkshop/RDA%20training%20materials/LC%20RDA%20Training/FRBR_Module%203_FRBR%20&%20RDA%20&%20MARC/FRBR%20%20RDA%20%20MARC_studentversion

Dublin Core is another metadata schema known for its extensive integration and use. It's proposed capability to describe almost any information object is the basis of its widespread appeal for collections. The element set of Dublin Core is grounded on the lowest common denominator principle, attempting to capture the "core" attributes to describe a work. This is established through the fifteen main elements of the structure, all of which are optional and repeatable. The Dublin Core Abstract Resource Model (Figure 2) displays how descriptions are modeled; very simply, the resource is described using a property-value pair, meaning, the property is the aspect being described (eg. title, creator, etc.) and the value is the word(s) assigned to that property to uniquely identify the property (eg. the name of the title, the name of the creator, etc.) The "resource" in the data model could be anything, including the work, expression, manifestation, or item. Beyond the fifteen elements, further details of object description are possible through the use of element attributes, which can establish relationships and context. Two significant features of Dublin Core concerning this analysis are its potential for wide application and its simplicity.

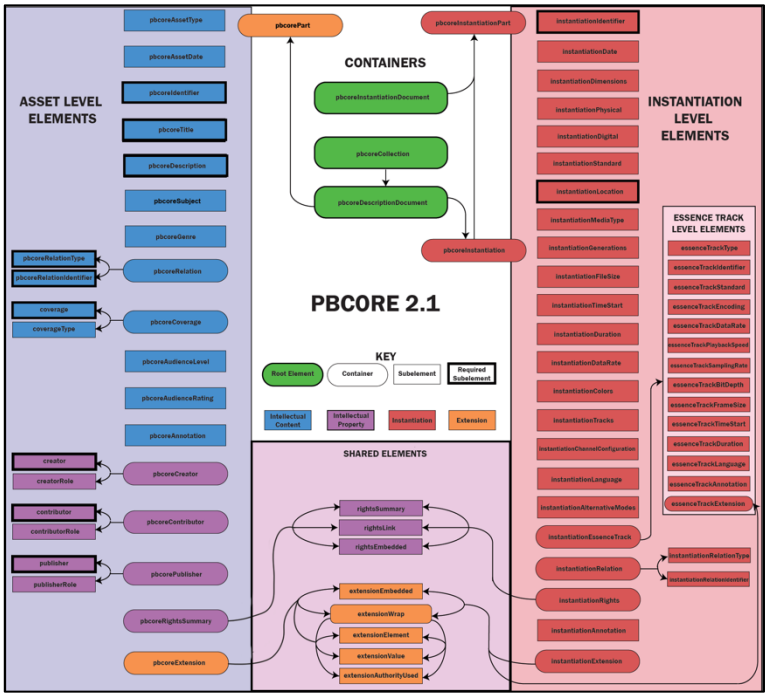
Figure 2. **Dublin Core Abstract Resource Model**



Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnston, Tom Baker, "DCMI Abstract Model," *Dublin Core Metadata Initiative*, June 4, 2007, https://www.dublincore.org/specifications/dublin-core/abstract-model/.

Looking towards metadata schemas created for specific formats, PBCore was deliberately developed for describing audiovisual materials. It is similar to Dublin Core, as it was derived from the standard, but PBCore was more purposely designed to include descriptive and technical fields for audiovisual material description. PBCore was originally created to give public broadcasting organizations the ability to manage the metadata of their media, however, it has been adopted more generally by moving image archives and libraries. It may be accompanied by other metadata standards for additional needs through extension capabilities, though at the basic level, it is focused on descriptive and technical metadata at the asset and instantiation level, which could also be thought of as the manifestation and item level of the FRBR model (see figure 3).[2]

Figure 3. **PBCore Data Model**

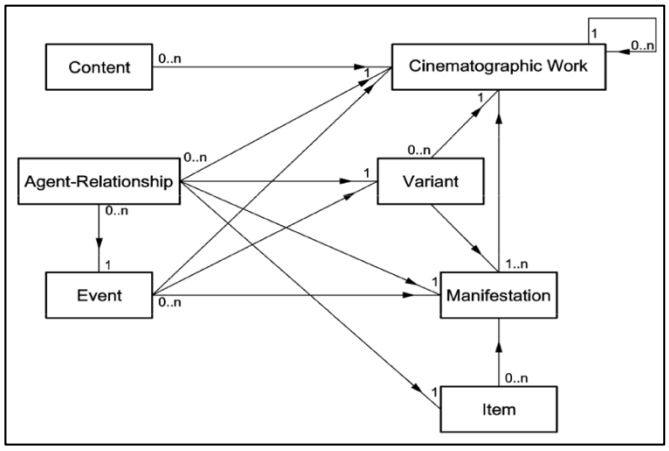


"PBCore Data Model Visualization," *PBCore*, accessed December 16, 2020, https://pbcore.org/data-model

The last metadata schema that will be discussed is the EN 15907, which is a European standard that structures the description of cinematographic works based on its primary entities, contextual entities, varying element types, and relationships. The primary entities of this standard align well with the FRBR data model, as EN 15907 similarly uses the entities such as cinematographic work, variant,

[2] "Frequently Asked Questions," *PBCore*, accessed December 16, 2020, https://pbcore.org/faqs.

Figure 4. **EN 15907 Data Model**



"Elements of the Data Model," *filmstandards.org*, accessed December 16, 2020, http://www.filmstandards.org/fsc/index.php?title=EN_15907

manifestation, and item, with the extension of the entities of content, agent, and event to further describe and contextualize. It is useful that on all of these four levels, agent-relationships can be assigned, and for three of the levels, (all excluding the item level) events can be attached, allowing for a data model more suitable for the nature of moving images. In this way, while the FRBR model can conceptualize description for potentially any information object in the universe, EN 15907's model can only be used for moving image description. Therefore, not only was the schema designed for moving images, but its use is exclusive to them as well.

Evaluating these metadata schemas to expose the varying levels of effectiveness will be done using the International Federation of Library Association's (IFLA) General Principles proposed in the *Statement of International Cataloging Principles* published in 2009.[3] The annotated listing of the nine general principles is as follows:

1. Convenience to the user: *How much does the schema consider the user's experience?*

   IFLA states this principle is to be most prioritized, as providing ease of use will likely be the most impactful for the user's experience with the catalog. It relates to other cataloging and metadata guidelines, such as those stated in Gilliland's work,[4] but also supports the user tasks that FRBR defines as

[3] "Statement of International Cataloguing Principles," *The International Federation of Library Associations and Institutions*, 2009, https://www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf.
[4] Anne Gilliland, "Setting the Stage."

essential abilities in a catalog—finding, identifying, selecting, and obtaining resources.[5] Convenience to the user during these research tasks is an essential goal of proper cataloging.

2. Common Usage: *Does the schema prioritize common knowledge and forms of use?*

   This principle also relates to Gilliland's work, when they state, "metadata presents some of the information that an information professional might have provided in a traditional, in-person reference or research setting."[6] The schema should outline metadata that is typically of interest to researchers. However, while this principle may be intended to support the broadest range of users, it can also leave room for faulty assumptions about who the users are, what they are interested in, and what "common usage" looks like, especially for a diverse user base. For that reason, this second principle should be considered carefully from many different perspectives.

3. Representation: *Does the schema allow for the resource to be described based on how it originally appears on the resource itself?*

   Representing a resource in the catalog is crucial for its discovery and recognizability. Therefore, this principle aligns with the FRBR promoted user tasks of finding and identifying.

4. Accuracy: *Does the schema require a precise description to achieve accuracy?*

   Representation of the resource is only of value if that representation is accurate. Once again drawing upon Gilliland's work, the ability of metadata

[5] IFLA Study Group on the Functional Requirements of Bibliographic Records, "Functional Requirements of Bibliographic Records: Final Report," *International Federation of Library Associations and Institutions*, September 1997, http://www.ifla.org/VII/s13/frbr/frbr.pdf.
[6] Anne Gilliland, "Setting the Stage."

to "certify the authenticity and degree of completeness of the content"[7] is essential to accurately represent a resource. Achieving this principle in the catalog could be demonstrated by the schema's allowance of precise elements, sub-elements, or some other form of faceted details.

5. Sufficiency and necessity: *Does the schema only require indispensable elements?*

   This principle is interesting in that it recommends only providing data elements "that are required to fulfill user tasks and are essential to uniquely identify an entity."[8] Of course, the inclusion of irrelevant data should be avoided, but also data that is too specific or too broad to the point where it doesn't assist in the finding, identifying, selecting, or obtaining a resource should be prevented.

6. Significance: *Are the elements required by the schema significant to the description?*

   IFLA explains this principle as "data elements should be bibliographically significant."[9] This is important for facilitating FRBR's user tasks and also acts as a counterbalance to the previous principle of accuracy. Data elements should be detailed and precise, but not so specific to the point that they are insignificant to the representation of the resource.

7. Economy: *Does the schema allow for efficiency and clear objectives?*

   The benefit of the cataloger is a key consideration in this principle. When there are multiple possibilities, IFLA states, "preference should be given to the way that best furthers overall economy (i.e., the least cost or the simplest approach)."[10] This may lead to benefitting the user as well. One criticism of

[7] Ibid.
[8] "Statement of International Cataloguing Principles," *IFLA.*
[9] Ibid.
[10] "Statement of International Cataloguing Principles," *IFLA.*

this principle is that it may contradict other IFLA principles that prioritize dedicated time and effort over efficiency (for example, ensuring accuracy, necessity, or consistency will likely not be a simple process, but it still included as a recommended principle because of its importance).

8. Consistency and Standardization: *How does the consistency of the schema affect description and access?*

   Although metadata schemas are standards themselves, their consistency will likely vary across their application. For example, some schemas allow for optional and voluntary element fields. This principle can guide analysis of how the level of flexibility of the schema's rules play a role in effective description and access.

9. Integration: *Does the schema rely on a shared set of requirements?*

   This principle can relate to the importance of interoperability between metadata schemas, as the adoption of common rules across schemas and repositories can benefit the mapping of their metadata. Integration of a common set of guidelines eventually allows for interoperability that results in further access to materials.

While this set of general principles has its imperfections, like the assumption of common use, the contradictory guidelines, and the absence of any principles regarding the importance of equitable description or culturally inclusive description practices, these nine principles have guided international cataloging codes for decades.[11] For the following evaluation of MARC, Dublin Core, PBCore, and EN 15907, these guidelines can expose the advantages and disadvantages of their varied capabilities. It has been

[11] Ibid.

decided that this is a suitable criterion to address how these schemas determine success for cataloging moving images because of three primary reasons: 1) In general, the set of principles consider both the user and the cataloger. 2) The principles can be applied to online library catalogs. 3) The principles are general enough to encompass all types of materials, beyond textual works, to include moving image cataloging as well. These last two points were not covered in IFLA's original 1961 Statement of Principles; however, the 2009 revised edition has included them as useful and necessary considerations. As the Statement currently stands, these nine principles should be key concerns when assessing metadata schemas.

MARC

Measuring the MARC schema against the IFLA General Principles reveals its few strengths and its several weaknesses. It seems to perform well in the IFLA terms of representation and integration. Its ability to represent a moving image is based on its provided bibliographic fields, such as the general physical description fields (007, 300) that allow for motion picture, playing time (306), projection characteristics (345), video characteristics (346), and so on.[12] It typically relies on a common set of rules, such as data value and content standards (LCSH, LCGFT, etc.), to promote integration. However, that is about the extent of its strengths for moving image cataloging. Most notably, MARC's downfalls appear to be its lack of accuracy, economy, convenience, consistency, and standardization. This evaluation is based on personal use and observation, as well as a recently conducted interview with UCLA Film and Television Cataloger, Amanda Mack.

[12] "MARC 21 Format for Bibliographic Data," *Library of Congress Network Development and MARC Standards Office*, December 2020, https://www.loc.gov/marc/bibliographic/.

For Mack, MARC may get the job done, but there are some clear downfalls, such as its lack of accuracy and economy, in particular.[13] This is likely a result of MARC not being a moving image specific schema. For example, while the MARC bibliographic 511 field is designated as the "Participant or Performer Note," the UCLA Film and TV archive has locally repurposed that field to be more of a "Cast Note." Moreover, the 250 field, known as the "Edition Statement" in MARC, actually functions as the "Version Note" in UCLA's moving image archive catalog. This absence of adequate terminology for moving image cataloging is just one aspect of inaccuracy. There are also limitations of the fixed fields, as Mack expressed the frustration of not being able to properly code a 4 ¾ video format because it is not provided as an option. So, while the representation of moving images may be possible, the most accurate representation is not optimized in a MARC record. There are also issues regarding its convenience to the user, just based on the high number of encoded elements that require specialized knowledge to comprehend.

Dublin Core

In contrast to MARC, Dublin Core offers a clearer demonstration of the IFLA General Principles in its schema. Convenience to the user, common usage, economy, sufficiency and necessity appear to be advantages of the schema, based on its plain text elements, simplicity, and the effort to primarily contain the core fundamental metadata. It's simplicity also allows it to be flexible, extensible, and compatible, allowing for a higher degree of integration.[14] The simplicity of Dublin Core has been successful in some of these regards; however, its simplicity can come at a cost. As discussed by Julie

[13] Amanda Mack (Cataloger, UCLA Film and Television Archive), interview by Jessica Craig, December 7, 2020.
[14] Julie Weagley, Ellen Gelches, and Jung-Ran Park, "Interoperability and Metadata Quality in Digital Video Repositories: A Study of Dublin Core," *Journal of Library Metadata* 10, no. 1 (2010): 37–57.

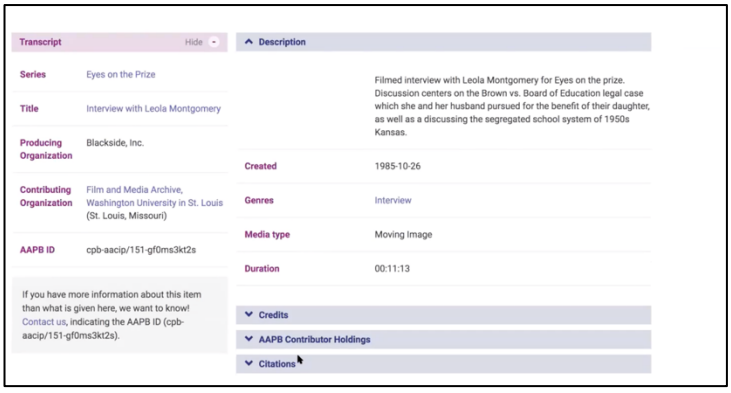
Weagley, Ellen Gelches, and Jung-Ran Park, Dublin Core's implementation across repositories can reveal some varying levels of completeness, accuracy, consistency, use of controlled vocabularies, and interoperability. For example, in regard to the completeness, they state: "Of the fifteen elements suggested by Dublin Core, only Title is supplied across the repositories 100% of the time. Following Title are Description (99%), Date (96%), Identifier (83%), Type (83%), and Relation (83%) … Four elements are collected at less than 50%, Creator, Contributor, Source, and Coverage." Not only does this reveal that Dublin Core records often don't use all 15 elements, but the consistency of which elements also differs. Another major issue, mentioned by Weagley, Welches, and Park, is the semantic vagueness of some Dublin Core elements, for example, Format, Type, Contributor, and Creator. This ambiguity can be a major downfall for integration. However, as they conclude, while the Dublin Core schema shows low consistency across repositories, it is actually quite high within them. Thus, the scope of the effectiveness of Dublin Core should be completed with this in mind. Overall, Dublin Core seems to reveal more advantages than MARC, while still not being most capable for moving image cataloging.

PBCore

As mentioned, PBCore was derived from Dublin Core to offer moving image catalogers a more appropriate schema for audiovisual formats. Its specific application promotes its convenience to the user, use of common terms, bibliographical significant elements, and economy. It's encoded in XML on the back end, which allows for more easily shared collection metadata and integration. As Rebecca Fraimow expressed, the XML data format allows for the inclusion of complex technical and descriptive metadata without bombarding the front-end user experience with information they

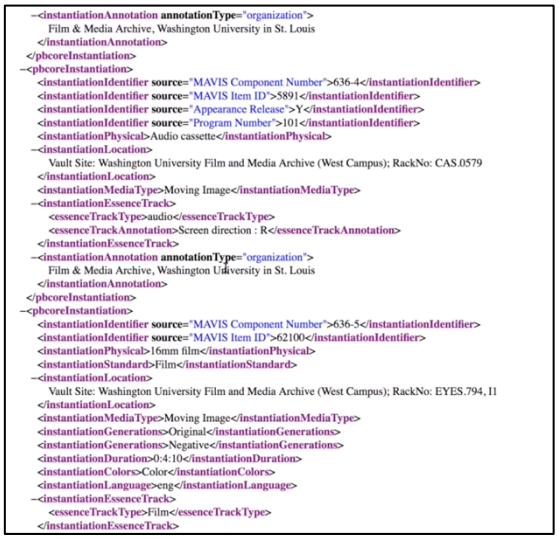
may not be interested in.[15] However, this effort towards convenience brings up questions of common use. In the American Archive of Public Broadcasting, much of the technical metadata found in the XML file is not all displayed on the front end, since it is assumed to be typically uninteresting (see figure 5). The encoded information still exists, but it is not made readily available to the user. If the inclusion of technical metadata is one of the reasons for PBCore's creation in the first place, perhaps it should be more accessible. The decision to exclude some of the technical metadata may align with the common use principle because perhaps, the technical information truly isn't commonly used. But the decision is backed by assumptions that should be explored. Largely, PBCore is a useful schema for moving image catalogs, as it considers the user, includes accurate moving image terminology, allows for significant elements and optimal representation, and promotes integration.

Figure 5. **American Archive Front Plain Text Record**



Front-end PBCore record displaying series, title, producing organization, contributing organization, a description, creation date, genre, media type, duration, credits, holdings, citations, in plain-text readable format. Screenshot excerpt from Rebecca Fraimow's discussion with May Haduong, November 2020.

Figure 5. **American Archive XML Record**



Back-end PBCore record expressed in XML encoding. This record holds more technical metadata not seen on the front-end of the same record. Screenshot excerpt from Rebecca Fraimow's discussion with May Haduong, November 2020.

[15] Rebecca Fraimow (Archivist, GBH), in discussion with May Haduong, December 2020.

EN 15907

Lastly, a look at how the IFLA's General Principles function in the EN 15907 metadata schema. There were some specific strengths called out by the basic structure and design. Like PBCore, EN 15907's specific use for moving image cataloging heavily benefits it's the effectiveness of its description and access to the format. In particular, access through interoperability. For example, while EN 15907's simpler counterpart schema 15944 is similar to Dublin Core but for moving images, 15907 is a more comprehensive schema, as Ronny Lowey explains, "designed to provide a pathway towards increased interoperability, both among film databases, and between these and other information systems"[16] Therefore, EN 15907 aligns well with the IFLA General Principle of integration, common use, and consistency. These features have been demonstrated through its adoption for union catalogs, such as the filmarchives-online.edu and europeanfilmgateway.eu.[17] Further research did not prove any disadvantages of the schema, although the limited amount of literature regarding this topic may be a reason for this. Generally, EN 15907 appears to be an optimal metadata schema for moving image cataloging, especially in regard to enhanced integration and interoperability.

Although the practice of cataloging is full of standards and guidelines, the process and result of cataloging moving images can look very different in and across repositories. The IFLA General Principles have become a useful measure to assess cataloging practices in general, as it can reveal the varying levels of effective description

[16] Ronny Loewy, "CEN Standards for Metadata About Cinematographic Works," June 2010, *Duetsches Filminstitut*, http://filmstandards.org/media/cen-cws-synop-2010-06a.pdf.
[17] "Metadata Management in Film Archives: Putting the 'Cinematographic Works Standard' EN 15907 to use and introducing the new FIAF Cataloguing Manual," *International Federation of Film Archives*, accessed December 2020, https://www.fiafnet.org/pages/Training/Metadata-Management-in-Film-Archives.html.

and access. For the four metadata schemas discussed in this paper, the IFLA principles were seen demonstrated in various ways. For MARC, although it could represent moving images sufficiently, its major criticisms included a lack of accuracy, economy, and convenience. With Dublin Core, its simplicity aided its convenience, common use elements, and economy, but cost accuracy and consistency. PBCore was much more effective all-around, with optimal convenience to the user, accurate terminology, and economy, although its application may bring up questions of common use. And finally, EN 15907, as a more recent development, demonstrates a high level of integration possibilities, with accuracy and representation also exhibited. In conclusion, it seems as though improvements of metadata schemas for moving images progress with time, as each new structure builds off the strengths and weaknesses of a previous one. With the adoption of principles such as IFLA's, it may be most successful, although, implementation will still likely vary depending on its use.

Bibliography

"FRBR: FRBR, RDA, and MARC." *Library of Congress Cooperative and Instructional Programs Division*. September 2012. https://www.loc.gov/catworkshop/RDA%20training%20materials/LC%20RDA%20Training/FRBR_Module%203_FRBR%20&%20RDA%20&%20MARC/FRBR%20%20RDA %20%20MARC_studentversion_20120818.pdf.

Gilliland, Anne J. "Setting the Stage." In *Introduction to Metadata*, edited by Murtha Baca. 3rd ed. Los Angeles: Getty Publications, 2016. http://www.getty.edu/publications/intrometadata/setting-the-stage/.

IFLA Study Group on the Functional Requirements of Bibliographic Records. "Functional Requirements of Bibliographic Records: Final Report." *International Federation of Library Associations and Institutions*. September 1997. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

"MARC 21 Format for Bibliographic Data." *Library of Congress Network Development and MARC Standards Office*. December 2020. https://www.loc.gov/marc/bibliographic/.

"PBCore Data Model Visualization." *PBCore*. Accessed December 16, 2020. https://pbcore.org/data-model.

"Frequently Asked Questions." *PBCore*. Accessed December 16, 2020. https://pbcore.org/faqs.

"Statement of International Cataloguing Principles." *The International Federation of Library Associations and Institutions*. 2009. https://www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf.

Weagley, Julie, Ellen Gelches, and Jung-Ran Park. "Interoperability and Metadata Quality in Digital Video Repositories: A Study of Dublin Core." *Journal of Library Metadata* 10, no. 1 (2010): 37–57.

LIST OF COURSEWORK

Course Number	Course Title	Instructor
Fall 2019		
IS 211	Artifacts and Cultures	Johanna Drucker
IS 260	Description and Access	Gregory Leazer
IS 432	Issues and Problems Preserving Cultural Heritage	Ellen Pearlstein
Winter 2020		
IS 270	Computer Systems and Infrastructures	Miriam Posner
IS 461	Descriptive Cataloging	Luis Mendes
IS 271	Human Computer Interaction	Leah Lievrouw
Spring 2020		
IS 212	Values and Communities in Information Professions	Ramesh Srinivasan
IS 462	Subject Cataloging and Classification	Luis Mendes
IS 464	Metadata	Melissa Gill
Summer 2020		
DH 101	Introduction to Digital Humanities	Ashley Sanders Garcia
COMM 140	Theory of Persuasive Communication	Michael Suman
Fall 2020		
IS 214	Informatics	Ramesh Srinivasan
IS 289	Museums in the Digital Age	Miriam Posner
IS 289	Media Description and Access	May Haduong
IS 498	Internship	Dee Winn
Winter 2021		
IS 438	Archival Description and Access	Jonathan Furner
IS 280	Social Science Research Methods	Leah Lievrouw
IS 400	Professional Development and Portfolio Design	Safiya Noble & Gregory Leazer
IS 239	Letterpress Lab	Johanna Drucker
IS 498	Internship	Dee Winn
Spring 2021 (in-progress)		
IS 439	Special Collections	Robert Montoya
DH 299	Designing the User-Centered Art Archive	Kathy Carbone
IS 498	Internship	Dee Winn

ADVISING HISTORY

Professor Miriam Posner has been my advisor since Winter 2021. My advisor changed from Professor Michelle Caswell to Professor Posner to reflect to my growing interests in digital humanities, which is one of Professor Posner's specializations. Throughout my first year and during the Fall of 2021, I formally met with Professor Caswell at least once per quarter. She supported my academic and professional interests by reviewing my course enrollment choices, recommending journal articles, professional conferences, scholarships relevant to my professional endeavors, and acting as a reliable source of insight and leadership in the IS department. I have similarly consulted with Professor Posner regarding my MLIS courses, progress on my issue paper, MLIS portfolio, and the Digital Humanities Graduate Certificate. Professor Posner has provided me with active mentorship since becoming my advisor in 2021. With both advisors, there were numerous emails exchanged in addition to our formal meetings. My scheduled meetings with both Professor Caswell and Professor Posner are documented below:

Professor Michelle Caswell

Introductory Meeting
September 25, 2019

Fall 2019 Meeting
October 10, 2019

Winter 2020 Meeting
January 14, 2020

Spring 2020 Meeting
May 13, 2020

Fall 2020 Meeting
October 2, 2020

Professor Miriam Posner

Winter 2021 Meetings
January 27, 2021
February 11, 2021

Spring 2021 Meeting
March 31, 2021