Information Services & Use 41 (2021) 53–60
DOI 10.3233/ISU-210117
IOS Press

Creating return on investment for large-scale metadata creation

Michelle Urberg*
Independent Consultant, Seattle, WA, USA

Abstract. The scholarly communications industry is turning its attention to large-scale metadata creation for enhancing discovery of content. Algorithms used to train machine learning are powerful, but need to be used carefully. Several ethical and technological challenges need to be faced head-on to use of machine learning without exacerbating bias, racism, and discrimination. This article highlights the specific needs of humanities research to address historical bias and curtail algorithmic bias in creating metadata for machine learning. It also argues that the return on investment for large-scale metadata creation begins with building transparency into metadata creation and handling.

Keywords: Metadata, machine learning, discovery, historical bias, algorithmic bias, humanities research

1. Introduction

Metadata is all about context. Key terms, titles, identifiers, and other pieces of information help to describe, classify, and facilitate the discovery of content. It is hard to comprehend the impact of any one piece of metadata, but easy to feel the impact of thousands or millions of pieces of metadata in a search experience. Metadata can have great value for end users: it fuels research and the production of new ideas, and – if done well – can create links across domains or surface previously hidden information. When done poorly or haphazardly, content is, at best hidden and at worst recast in unsavory or dangerous ways, which are compounded at scale.

In the discussion that follows, I explore some of the challenges to creating robust metadata at scale with machine learning algorithms. I also elaborate on some concrete effects of bias that arise from poorly constructed metadata or the poor use of metadata. I then argue that return on investment for large-scale metadata creation begins with building transparency into metadata creation and handling.

The metadata of interest to this article is created for scholarly outputs related to humanities research. Metadata for STEM scholarly outputs is, comparatively much talked about, including in presentations from the first NISO Plus conference [1–3]. Humanities datasets and research objects have received far less attention by contrast. Thankfully, however, the pitfalls of machine learning discovered by scientific publishers serve as learning models for humanities publishers. The time has come to explore what humanities projects require for dataset analysis and what humanities publishers need from machine learning to enhance their metadata. I present six challenges below that should be at the forefront of any discussion about creating return on investment for metadata, with attention to the particular needs of humanities scholarly outputs.

*E-mail: maurberg@gmail.com.

2. Challenges

Challenge 1: *Creating reusable metadata for special data sets*

The creation of metadata for humanities research invites certain challenges with qualitative data sets that are not often present with scientific research. First, data can include images, video, art works, musical scores, interviews, and other qualitative materials. Second, these objects often lack robust standards to describe them (e.g., interviews) or lack technology to make descriptive metadata creation scalable (e.g., manuscripts, musical scores, art works). Third, few places exist to deposit these types of datasets.

I personally know what it is like to wish for better metadata creation mechanisms to support a large-scale humanities project. I spent many months photographing fourteenth-, fifteenth-, and sixteenth-century manuscripts and then later cataloging their contents. I meticulously tracked how women were depicted iconographically or transcribed music to document changes (or possibly mistakes!) in melody or muiscal text. While I loved getting my hands dirty with these materials, I did not know of an ideal place to deposit data from my dissertation when I finished all making all of these notes, meaning that I had no way to standardize the data for deposit into an existing search platform [4]. Moreover, I had very little incentive to create metadata or a database of the music and image that I studied. The value of my dissertation lay in its interpretation of the musical culture, rather than in the cataloging of image and music. But absent a repository or descriptive metadata, the value of my images and transcriptions extended only as far as my dissertation. As an end user, metadata or query-able datasets would have been of enormous value both to my dissertation and for the broader community of medieval musicologists.

Challenge 2: *Creating metadata for machine learning*

The creation of metadata useable in machine learning begins with training algorithms how to work with information – in this case metadata describing scholarly outputs – to improve the search and discovery of content. To train an algorithm to make humanly-acceptable decisions about sets of information, the process begins with selecting sets of metadata related to the metadata associated with production content. Ideally, training sets represent content rich and correctly-formatted information including, but not limited to, relevant words, phrases, and correct domain-specific idioms. When a schema, standard, Document Type Definition (DTD), or taxonomy applies, it should be used. If the training set does not structurally, thematically, or idiomatically relate to the production data set, the training of the algorithm will be less effective [5]. In short, to produce trustworthy macine-generated tagging or content analysis, the machine doing the algorithmic learning needs to have relevant and trustworthy inputs.

Challenge 3: *Creating vocabulary that acknowledges historical bias*

The creation of vocabularies (e.g., taxonomies and terms used for tagging) to help train a machine learning algorithm requires subject matter expertise, as well as the acknowledgement of the historical contexts in which the content was created. Historical bias is the result of a policy, practice, or other decisions that have larger reaching consequences than was intended by the policy, practice, or decision. It is an unintended byproduct. Vocabulary that is important for training machine learning algorithms can be negatively influenced by the bias that has accrued to domain specific terms. This may mean that a concept is either overly emphasized, hidden, or misrepresented through the training process. In other words, vocabulary needs to be chosen carefully and used with human intervention to prevent undesirable consequences arising with using algorithms to automatically generate metadata.

Building a useful set of terms to use as a training vocabulary can be fraught with a number of challenges. Commonly used terms that could serve as subject tags may not adequately describe a concept or existing

terms may not translate across domains. The word *rolig* in Norwegian, for example, means calm, but in Swedish it means fun. Or, as the *Keywords* series by New York University Press demonstrates, terms such as "gender" may either restrict a description or do not neatly fit for a given field of study [6]. Sandra K. Soto describes it in her book *Keywords for Latina/o Studies*: "Gender" is difficult: "Like the terms with which it most often travels ("race", "sex", and "sexuality"), gender is a complex and contested concept that, although used quite widely and more and more frequently in both academic and nonacademic contexts, means significantly different things to different people and across different institutional locations" [7].

Moreover, well-established vocabularies such as the Library of Congress Subject Headings (LCSH) or the Library of Congress Genre Form Terms (LCGFT) have been created largely in the United States and reflect an Americo-centric perspective on subject classification, which has been discussed in library circles for years [8,9]. The goal of LCSH and LCGFT has always been to provide reusable and shareable classification terms that make library catalogs, and now library discovery systems, work. While they do generally "work" for library discovery in English-speaking/reading contexts, they do not serve global needs nor is the classification taxonomy nuanced enough across the vocabularies to meet the needs of many descriptions (e.g. consider the contentious discussion about "illegal aliens" and "undocumented immigrants").

All of the challenges with vocabulary construction are also long-standing problems for description of humanities scholarship that make it hard to tag content with accurate, equitable, and discovery-friendly metadata. Historical bias makes metadata training sets ineffective at best and harmful at worst.

Challenge 4: *Acknowledging systemic problems with metadata*

The creation of metadata for machine learning should acknowledge that systemic bias is real and needs to be exposed. In the IEEE Spectrum blog from April 15, 2019, Oscar Schwartz chronicles the systematic algorithmic bias that was built into the St. George's Hospital Medical School application process during the early 1980s [10]. St. George's disproportionately weighted name and place of birth to reject certain people from the initial screening process (and therefore rejecting them from the school). Based on name and place of birth, prospective students were labeled as either "Caucasian" or "non-Caucasian," with points deducted from the application for non-Caucasian applicants. Fewer points meant that the candidate was less likely to be chosen to interview for a position to study medicine. In aggregate, the effect of this one term led to intentional racial and gender discrimination and ultimately the loss of diversity gains made in the later 1970s at the school [11]. Meanwhile, this bias enacted an institutional block against those applicants who were rejected by an algorithm, thus changing career trajectories for a significant number of candidates.

While this discrimination case was widely covered in the mid-1980s and a commission investigated the issue, the inherent problems of this kind of bias have not been resolved in other programs, institutions, or governments and certainly not in internet searching algorithms [12]**.** The algorithmic learning that privileges one term or a set of terms to restrict or enhance discovery of a scholar, an institution, or work from a particular region is readily reproducible in any system that collates information for end-user discovery, including library discovery layers, content aggregators, and in pre-print servers. If the scholarly communications industry seeks to prevent this type of systemic bias from further infiltrating machine learning algorithms, it must openly acknowledge when problems arise in platforms using search trained through machine learning programs.

Challenge 5: *Updating problematic algorithms*

The creation of metadata for Machine Learning requires constant vigilance to fix or improve existing algorithms that produce search results. Safiya Umoja Noble, in *Algorithms of Oppression: How Search Engines Reinforce Racism*, for example, pushes Artstor (see https://artstor.org) to be more vigilant about

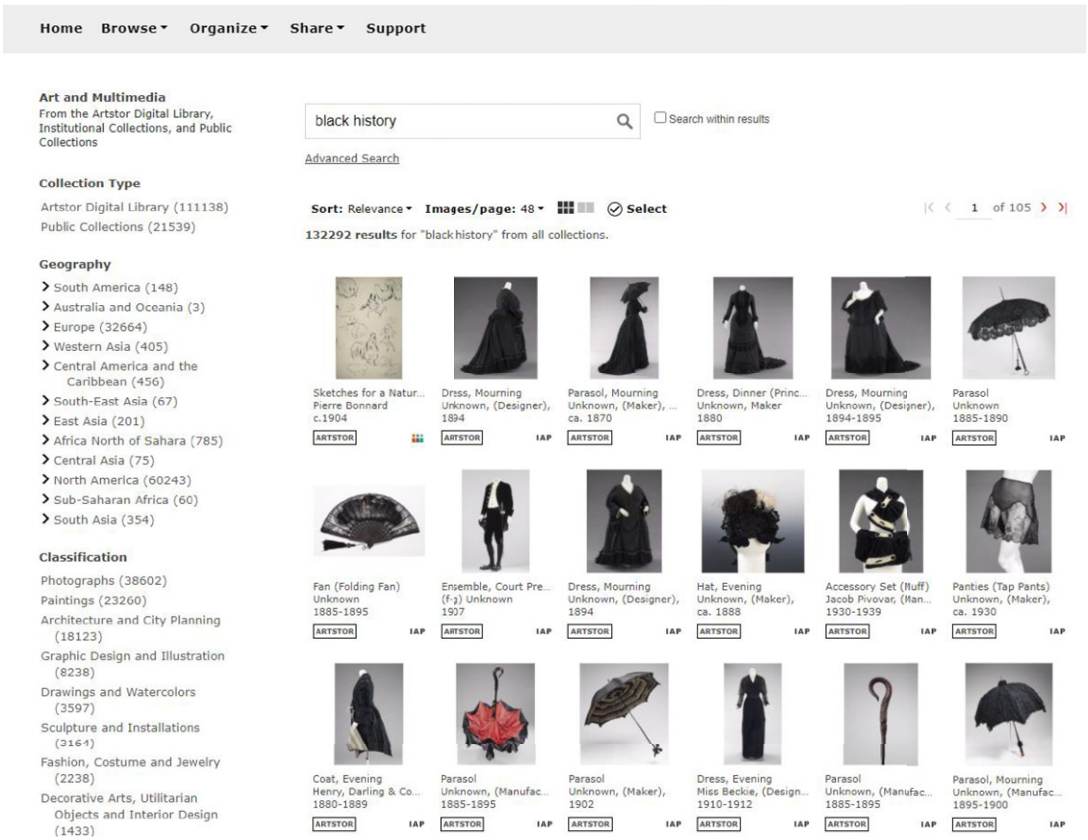

Fig. 1. March 2021 Artstor search on term "black history".

improving the search results for "black history", "African American stereotype", and "racism". When Noble conducted these searches in 2016 for publication in her book, the first results for "black history" were a number of European and white artists; the first result for "African American stereotype" is *On to Liberty,* a German painter Theodore Kauffman; and the term "racism" included a satirical piece entitled *Rent-a-Negro* [13]. The metadata attached to these images are framed by Noble as the "white racial gaze on information" and "a result of the investment of the [librarian] profession in colorblind ideology" [13]. From an end-user perspective, however, they are confusing at best and offensive at worst.

In March 2021, I revisited Noble's searches in Artstor. "Black history" (Fig. 1 below) now returned black clothing, from the 19th and 20th century in the first page of hits, but the sheer quantity of results has diluted the usefulness of this particular search.

"African American stereotype", (Fig. 2 below) by comparison, returns fewer results, but arguably inscribes stereotypes associated with slavery in art. Is this a worse type of misidentification than the results Noble found in 2016?

The results for "racism" (Fig. 3 below) are equally problematic, featuring images of riots and works of art using the term "racism" in their name.

The tagging and metadata associated with culturally-charged images need continual tuning. I would like to draw your attention to one key difference in the searches between 2016 and today, in 2021: Artstor has demonstrably improved the user experience to include a multitude of facets absent in 2016. While an

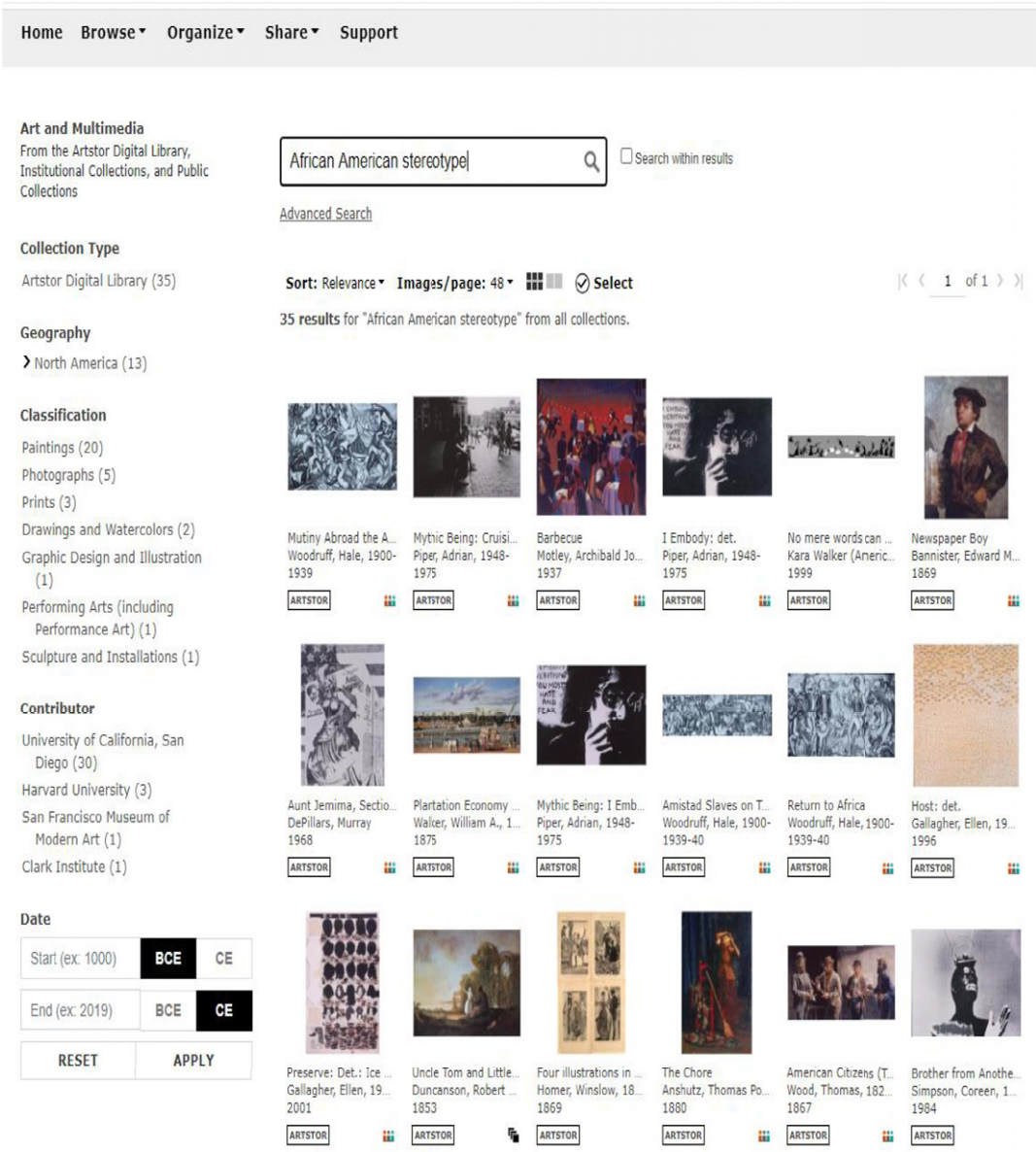

Fig. 2. March 2021 Artstor search on term “African American stereotype”.

abundance of limiting options is known to overwhelm users, in this case with the term “racism”, we see acknowledgement of how different media from around the world depict racism. These facets represent a lot of work toward acknowledging issues with algorithmic misidentification in the Artstor platform. Metadata is always a work in progress.

Challenge 6: *The allure of automation*

The creation of metadata for machine learning requires a respect for the power of automation, which can confound systemic bias of many types. Nevertheless, the promise of automation is alluring to business

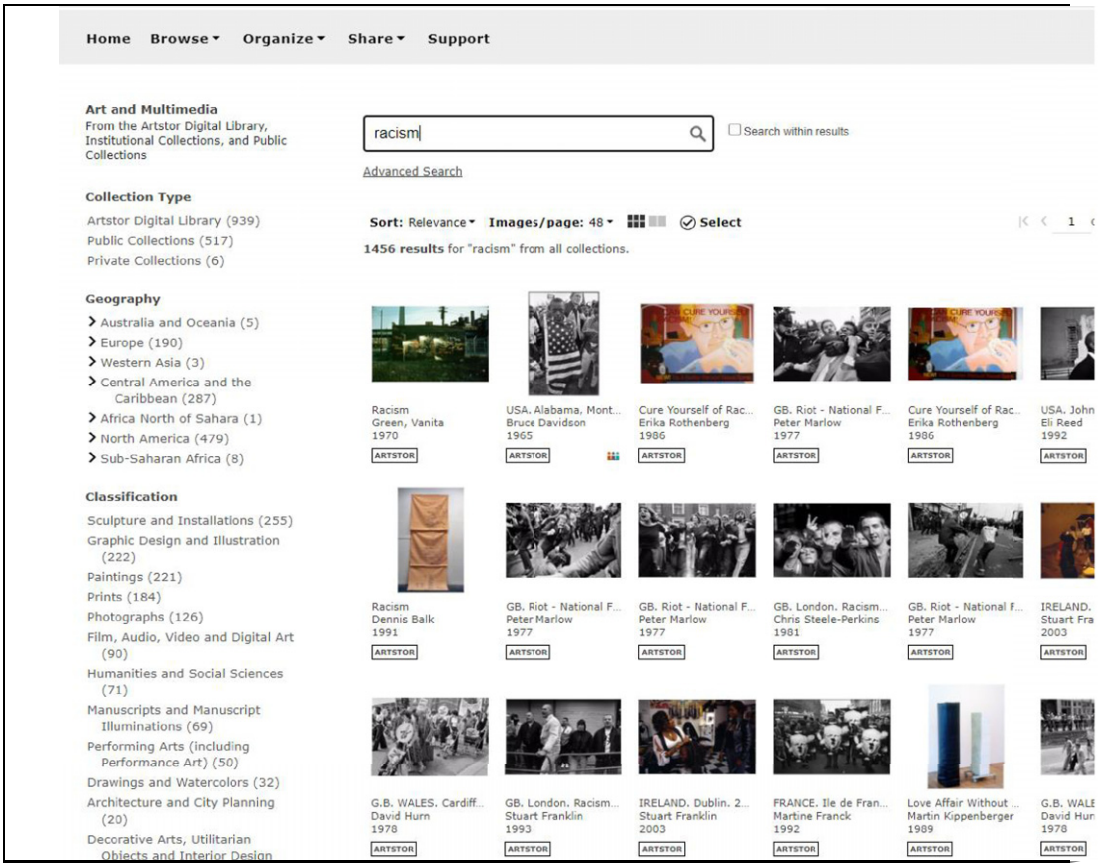

Fig. 3. March 2021 Artstor search on term “racism”.

needs: the more automated a process, ideally the less it affects the bottom line. Metadata is most attractive to business needs when it “just works”; unfortunately, it never “just works”. In order for metadata to be useful to end users and to produce a positive return on investment for publishers, it must be curated regularly. Metadata often requires more attention than it is given. It is no wonder that automation is a beacon of light when companies face the prospect of perpetual human-intensive metadata upkeep.

Automation, however, has the danger of compounding biases or problems in metadata. It can cause what David Lipsey, a leading digital asset management strategist calls “language decay”, a concept borrowed from linguistics that refers to a break in transmission of a language such that generations of speakers gradually lose the context, mood, tense, or morphology of that language. If controlled vocabularies are used in automated processes to tag content without proper context, it is certain that words will take on machine-generated meanings of their own. Language decay is a more insidious problem than refusing to remove outdated terms or naming conventions in controlled vocabularies. Language decay can compound cultural values and unconscious bias inherent in cataloging and metadata creation [14,15]. As both the St. George Hospital and Artstor examples show, automation can enhance racism that is already implicit in society. Bias in metadata must be addressed before processes are automated and automated processes must be carefully deployed to avoid the most dangerous types of ”language decay”.

3. Finding solutions

The creation of metadata for machine learning requires new solutions that may challenge existing business models in the scholarly communications industry. One way forward is to change how metadata is created and curated for training sets. Vocabularies are human inventions: if the vocabularies are or have become problematic, change the vocabularies (use related terms and synonyms to retain effective discovery). Likewise, subject matter expertise is important: lean into those who know a field or topic and look for vocabularies built to suit specific needs [16]. Subject matter experts (SMEs) can deeply influence the creation of metadata and vocabulary for the better: they can push back on the pitfalls of historical bias that are baked into LCSH, LCGFT, and other commonly-used vocabularies. SMEs can help solve the problem of inadequate or inappropriate training sets.

Another way forward is to invest in data governance models that confront historical bias head on. This could mean forcibly recasting how an algorithm interprets a given topic. It could mean publicly acknowledging the problems that are inherent in a discovery and search platform, as well as highlighting steps toward returning more inclusive, diverse, or equitable results. Likewise, it could mean welcoming and integrating user experiences for the improvement of search and discovery: if your product does not speak to the needs of your customers, why bother? In the spirit of NISO, I suggest that a Standard or Best Practice might be established to define ethical practices for using machine learning in scholarly publishing and in content discovery. Industry partners that developed JATS [17], KBART [18], and other metadata standards have made sharing possible in an agreed framework: why not for machine learning and its algorithms?

4. Creating return on investment

For machine Learning to effectively analyze humanities research, the metadata associated with the outputs needs to be carefully curated and the subject matter understood. To see return on investment for this approach in large-scale metadata creation, the perception of positive return will likely need to be calculated in terms of social good rather than in dollars and bottom lines. Just to be clear, I am writing here specifically to the danger of metadata for persons identifying as Black, Indigenous, People of Color and anyone else who considers themselves or their work to exist in intersectional spaces (gender, sexuality, race, disability, etc.). Metadata 20/20, an organization that I have helped support for several years, champions the cause of metadata as a tool by which we can save the world [19]. I whole heartedly believe metadata can save the world, but its power is such that it can and has systematically changed and destroyed lives. In more benign ways, word choice and algorithms affect how research is conducted, who is funded, or how conclusions are reached. It is time for the scholarly communications industry to tackle these challenges head on.

Acknowledgements

Thank you to Lettie Conrad for providing commentary on this draft. Thanks also to Verletta Kern for helping run the Artstor queries.

About the author

Michelle Urberg is an Independent Consultant. She has a PhD in Music History and an MS in Library and Information Science and is passionate about improving the scholarly communications life cycle. Her

work can be found at the Humanities Commons: https://hcommons.org/members/murberg/ and at ORCID: https://orcid.org/0000-0002-2748-8. Email: maurberg@gmail.com.

References

[1] J. Chabak, Artificial intelligence and machine learning 101, NISO Plus 2020, Available from: https://www.niso.org/niso-io/2020/03/niso-plus-artificial-intelligence-and-machine-learning-101, accessed September 23, 2021.
[2] B. Cody, AI and Machine Learning Presentation. Available from: https://www.niso.org/niso-io/2020/03/ai-and-machine-learning-presentation, accessed October 31, 2021.
[3] H. Wang, Data Discovery and Reuse – AI Solutions & the Human Factor. Available from: https://www.niso.org/niso-io/2020/03/data-discovery-and-reuse-ai-solutions-human-factor, accessed October 31, 2021.
[4] The one reliable platform to index medieval music is the Cantus Index. I used this database heavily, but it would have been a completely separate project prepare transcriptions for my dissertation. Cantus Index. Available from: http://cantusindex.org/, accessed September 23, 2021.
[5] H. Wang, Data Discovery and Reuse: AI Solutions & the Human Factor. 2020. Available from: https://doi.org/10.6084/m9.figshare.12081204.v1, accessed September 23, 2021.
[6] Keywords. Available from: https://keywords.nyupress.org/, accessed September 23, 2021.
[7] "Gender" search on Keywords. Available from: https://keywords.nyupress.org/search/?s=gender&cc_gs_sites%5B%5D=african-american-studies&cc_gs_sites%5B%5D=american-cultural-studies&cc_gs_sites%5B%5D=asian-american-studies&cc_gs_sites%5B%5D=childrens-literature&cc_gs_sites%5B%5D=comics-studies&cc_gs_sites%5B%5D=disability-studies&cc_gs_sites%5B%5D=environmental-studies&cc_gs_sites%5B%5D=latina-latino-studies&cc_gs_sites%5B%5D=media-studies&post_type=all, accessed September 23, 2021.
[8] T. Dankowksi, Removing Barriers to Indigenous Knowledge: IFLA session suggests ways to improve access to materials. 2016. Available from: https://americanlibrariesmagazine.org/blogs/the-scoop/removing-barriers-to-indigenous-knowledge/.
[9] Problem Library of Congress Subject Headings. Available from: https://cataloginglab.org/problem-lcsh/, accessed September 23, 2021.
[10] O. Schwartz, Untold History of AI: Algorithmic Bias was Born in the 1980: A medical school thought a computer program would make the admissions process fairer – but it did just the opposite, April 15, 2019. Available from: https://spectrum.ieee.org/untold-history-of-ai-the-birth-of-machine-bias, accessed September 23, 2021.
[11] S. Lowry and G. Macpherson, A blot on the profession, *British Medical Journal* (1988), 657–658.
[12] d. boyd, Questioning the legitimacy of data. Available from: https://niso.cadmoremedia.com/Title/cb.5f0733-cc96-4abc-a02f-121158d4a904, accessed October 31, 2021.
[13] S.U. Noble, *Algorithms of Oppression* 2018, p. 145.
[14] Language decay in metadata can be further complicated by the effects of regular use of search engines to extend our memory. Taylor & Francis. (16 August 2016). "Cognitive offloading: How the Internet is increasingly taking over human memory." *ScienceDaily*. Retrieved September 20, 2021. Available from: https://www.sciencedaily.com/releases/2016/08/160816085029.htm, accessed September 23, 2021. This blog references this article: B.C. Storm, S.M. Stone, A.S. Benjamin. (2016). "Using the Internet to access information inflates future use of the Internet to access other information." *Memory*. DOI: 10.1080/09658211.2016.1210171, accessed October 31, 2021.
[15] H. Schwartzman, "Google: An External Hard Drive for our Memory," November 26, 2019. Available from: https://web.colby.edu/cogblog/2019/11/26/the-google-effect-external-hard-drive/, accessed September 23, 2021.
[16] See for example Cataloging Lab's List of Alternative Vocabularies: https://cataloginglab.org/list-of-alternative-vocabularies/.
[17] H. Wang, Standardized Markup for Journal Articles Tag Suite (JATS), NISO. Available from: http://www.niso.org/standards-committees/jats, accessed October 31, 2021.
[18] Knowledge Bases and Related Tools, NISO, https://www.niso.org/standards-committees/kbart, accessed September 23, 2021.
[19] Metadata 20/20. Available from: https://metadata2020.org/, accessed October 31, 2021.