VRA Bulletin

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

Article 3

March 2019

Cataloging Today: Enlarging the Sphere

Susan J. Williams
Independent Consultant, williams.susanjane@gmail.com

Cataloging Today: Enlarging the Sphere

Abstract

Introductory comments from the guest editor are provided as a foreword to this special, themed *VRA Bulletin* issue on "Cataloging Today: Enlarging the Sphere." The editor explores cataloging practices in the light of the global semantic web and discusses the contributions to this issue.

Keywords

data catalog, Gartner reports, siloed data projects, monetization, commodification, LAMs, LOD, AI, DAMs

Author Bio & Acknowledgements

Susan Jane Williams is a 28-year member of the Visual Resources Association, and was its vice president in 1999-2000 (for the Chicago conference, and for San Francisco, the first solo conference). Following a career as a botanical illustrator, Williams graduated from Ohio State University with an MA in Art History, and began working as a slide and media librarian at Rochester Institute of Technology in 1990. Because Rochester is the home of Kodak, she was involved in some early and innovative digitization projects. Hired by Yale University, she was charged with building their first digital teaching collection. She won the DeLaurier Award from VRA in 2007 for developing a cataloging database tool, VCat, which utilizes VRA Core 4 and CCO and is capable of Core 4 XML, and now, RDF with LOD output. She has worked as an independent cataloger and consultant since 2007, serving the academic and commercial image vendor communities.

This perspectives is available in VRA Bulletin: https://online.vraweb.org/vrab/vol45/iss2/3

Cataloging Today: Enlarging the Sphere

Introduction

Since at least the first decade of the Visual Resources Association's founding, there has been discussion about how to enlarge the sphere of the mission and particularly, the membership, of the organization. The door has been open to members from the non-academic and commercial sectors for decades, in terms of the written mission statement.[1] However, a reason for those professionals to step across the threshold has been lacking; there has been little demonstrated overlap in our interests and practices, and our audiences and clients are seen as distinct.

The semantic web and 21st century cataloging practices may represent a significant change in a Venn diagram of our spheres of practice—this opportunity should be examined and a new dialog should perhaps begin. Our existing sphere is largely that of LAMs—libraries, archives and museums, plus non-library (i.e. department) based collections in higher education; in other words, the non-profit sector. Instead of focusing on the barriers that might arise when looking at individuals and their professional goals (as with past membership discussions), it might be more instructive to look anew at the larger environments for understanding both the barriers and the commonalities.

The For-Profit and Non-Profit Spheres

We are familiar with "data dictionaries," but the new tool in data management is the "data catalog," which seeks to cohesively manage and organize multiple data sets. Looking at a recent Gartner report on the topic, even a superficial scan of the online abstract can reveal commonalities, albeit couched in different prose.[2] Most are familiar with the Gartner reports for the education sector, but may not read the corresponding reports for the business sectors. The Gartner abstract maps the same concerns we face: managing "data sprawl" while trying to supply distributed data assets, compatibility problems with numerous vendor products and standards choices, and facing siloed data projects. This last point is phrased in an interesting way: "Siloed data catalog projects will limit the inventory of information assets and data monetization." LAMs have struggled with how we convey the value of collection creation and cataloging to our users and to our administrators; in the business world, they have a direct metric, i.e., monetization.

A recent *VRA Bulletin* article by Jasmine Burns seeks to make this very connection in the academic sphere.[3] One barrier that exists in the non-profit sector is, as she states, "maintaining core values and open access to information as a public good, while still participating in the market structures in which libraries and cultural heritage institutions are entrenched." Rather than direct 'monetization,' the emphasis here is placed on 'commodification,' which allows for a broader interpretation of markets, yet still is firmly rooted in concrete value and return on investment (of labor, among other things). Burns advocates for at least one solution, which might be applied currently; new, more balanced relationships and, most importantly, contracts between LAMs and commercial vendors in which in-house LAM labor is rewarded with direct or indirect revenue, and at the same time commercial exclusivity term-limited, with assets becoming open after a fixed period.

Connecting Technical Advances Between the Spheres

Cataloging and data management have been rapidly evolving towards the use of machine-driven and machine learning solutions for data creation, remediation and curation of data sets. As noted above, the creation of data catalogs is now practiced not only to aggregate data, but to link it. Some commercial applications were surveyed in the Gartner report, but the National Network of Libraries of Medicine also has members developing data catalogs.[4] Another point to ponder is the division between arts, humanities and the sciences, which further segregates our practices and knowledge. And, we should note more often that the semantic web is indeed a universal sphere, not one limited to education and LAMs. Linked Open Data (LOD) is being applied by Google and to products in the Amazon marketplace, which should clearly signal that LOD is not a passing fad. The U.S. educational sector, while boasting over $547 billion in annual endowments, is only part of the U.S. Gross Nation Product (GNP) of about $19 trillion (as of the third quarter of 2018).[5]

We are beginning to see the use of AI (Artificial Intelligence) in data applications. We are all familiar with the speech recognition that makes Apple's Siri, Microsoft's Cortana, Amazon's Alexa, and the Google Assistant able to process and respond to queries. In digital asset management systems (DAMs) specifically, AI use at this point breaks down to tools in a few areas, notably: batch comparison using contextual hints, pattern recognition and machine learning. In an online industry trade journal, Ralph Windsor points out a key issue in these techniques.[6] They all rely on learning from examination of initial data sets that have been properly cataloged by, as he amusingly puts it, the "metadata-virtuous," who have already been following best practices and policing the accuracy and validity of their metadata. This, and the fact that LOD also requires specific structure and the accurate application of that structure, suggests that information professionals in both the public and private spheres still have much to contribute. However, Stanford libraries are also exploring applications that can use what is termed "unsupervised" learning, "wherein a great deal of data is needed, but it need not be human-cultivated training sets."[7] Specific to machine learning involving images, it was noted in relation to a case study using images of cats, that this can require millions of images, which further suggests that librarians and collections developers also have a role here. However, it is worth noting that Amazon has developed jobs for "piecemeal tasks, such as labeling data" through its site Mechanical Turk (https://www.mturk.com/). Started in 2005, Mechanical Turk is favored by academic researchers for collecting survey and experimental data.[8]

This issue's article, "Automation and the Semantic Web: The Shifting Role of the Metadata Librarian," by Marlee Graser and Melissa Burel, is an excellent overview of the technologies mentioned above and more, including the practical daily use of them, even in smaller institutions today. The article is a must read for those who are wondering what technical skills they might cultivate now to meet the challenge of implementing automation in their workflow. The authors also make the point that while adding technical skills to one's resume is important, the evolution towards more programming and technical approaches also means that the work will become more, not less, collaborative. No one person will have the expertise necessary, and not only are metadata teams necessary, but so are networking skills and knowledge more broadly—enlarging the sphere of practice.

Connecting Practitioners

Amplifying the conclusions of the Graser and Burel article, the second article by Darcovich, Flynn and Li, "Born of Collaboration: The Evolution of Metadata Standards in an Aggregated Environment," provides a case study showing the need and a process to create or restructure metadata teams and administer them in this more technical and linked environment. The creation of new cross-departmental teams, guidelines and workflows at University of Illinois at Chicago (UIC) was done in the context of a focus on usability. Not only are the teams collaborative but there is designed workload sharing. The article also touches upon practical solutions in cross-collection data remediation and standardization.

Extending Standards and Connecting the Specialized Sphere

The third article, "Costume Core: Metadata for Historic Clothing," by Arden Kirkland, serves to illustrate the value of extending an existing metadata standard (VRA Core) for use with specialized material, as well as making sure the needed specialized terms for that area of study have a path to become incorporated in controlled vocabularies. Several related projects are outlined in this article, but one helps bring the special issue theme full circle through an interesting project (http://DressDiscover.org), to create a controlled thesaurus that can be searched visually, particularly by the non-specialist. The images are connected to AAT (the Art & Architecture Thesaurus) terms within the web application. As Kirkland notes, "there is potential also for this app to be used in reverse as a tool for constructing a search query based on visual choices, without having to know the correct vocabulary." Following the larger issues of LOD, machine learning and the semantic web, one can imagine how this project, generated from a specialized field of study, might scale up and not only connect specialized material to a broader audience, but help reinvent searches for all visual resources.

[1] Visual Resources Association, Mission Statement, http://vraweb.org/about/mission/

[2] Ehtisham Zaidi, et al. Data Catalogs Are the New Black in Data Management and Analytics, *Gartner*, 12/13/2017, https://www.gartner.com/doc/3837968/data-catalogs-new-black-data

[3] Jasmine E. Burns, "Information as Capital: The Commodification of Archives and Library Labor," *VRA Bulletin*: Vol. 45: Iss. 1, Article 9. Available at: https://online.vraweb.org/vrab/vol45/iss1/9

[4] National Network of Libraries of Medicine, https://nnlm.gov/data/thesaurus/data-catalog Accessed 2/26/2019

[5] National Center of Education Statistics, https://nces.ed.gov/fastfacts/display.asp?id=73 Accessed 2/26/2019 and Bureau of Economic Analysis, U.S. Department of Commerce. https://www.bea.gov/ Accessed 2/26/2019

[6] Ralph Windsor, "Artificial Intelligence and Metadata Cataloguing: Advice for Digital Asset Managers", *DAM News*, https://digitalassetmanagementnews.org/features/artificial-intelligence-metadata-cataloguing-advice-for-digital-asset-managers/ Accessed 2/26/2019

[7] Catherine Nicole Coleman, "Artificial intelligence and the library of the future revisited", *Digital Library Blog* post, Nov. 3, 2017 Available at: http://library.stanford.edu/blogs/digital-library-blog/2017/11/artificial-intelligence-and-library-future-revisited Accessed 2/26/2019

[8] Paris Martineau and Louise Matsakis, "Why It's Hard to Escape Amazon's Long Reach", *Wired Magazine*, 12/23/2018. Available at https://www.wired.com/story/why-hard-escape-amazons-long-reach/ Accessed 2/26/2019