VRA Bulletin

Volume 46
Issue 1 *Spring/Summer 2019*

Article 3

June 2019

Serving the Inquisitive User: A New Look at Tried and True Tagging (and some AI) Strategies

Lisa Gavell
Ithaka, lisa.gavell@ithaka.org

Follow this and additional works at: https://online.vraweb.org/vrab

 Part of the Art and Design Commons

Recommended Citation
Gavell, Lisa (2019) "Serving the Inquisitive User: A New Look at Tried and True Tagging (and some AI) Strategies," *VRA Bulletin*:Vol. 46: Iss. 1, Article 3.
Available at: https://online.vraweb.org/vrab/vol46/iss1/3

This Feature Articles is brought to you for free and open access by VRA Online. It has been accepted for inclusion in VRA Bulletin by an authorized editor of VRA Online.

Serving the Inquisitive User: A New Look at Tried and True Tagging (and some AI) Strategies

Abstract
Supported by user surveys, interviews, and analysis of usage metrics across collections, Ithaka is exploring two separate methods by which to attach subject terms to Artstor metadata. Future user surveys and interviews will seek the feedback and guidance of Artstor users.

Keywords
metadata, AI, Artificial Intelligence, tagging, Artstor

Author Bio & Acknowledgements
Lisa Gavell is Artstor's Senior Manager of Image Content. She manages metadata within the Artstor Digital Library. She holds an M.A. in American Studies from Columbia University and a B.A. from Simmons College in Art History.

This feature articles is available in VRA Bulletin: https://online.vraweb.org/vrab/vol46/iss1/3

INTRODUCTION

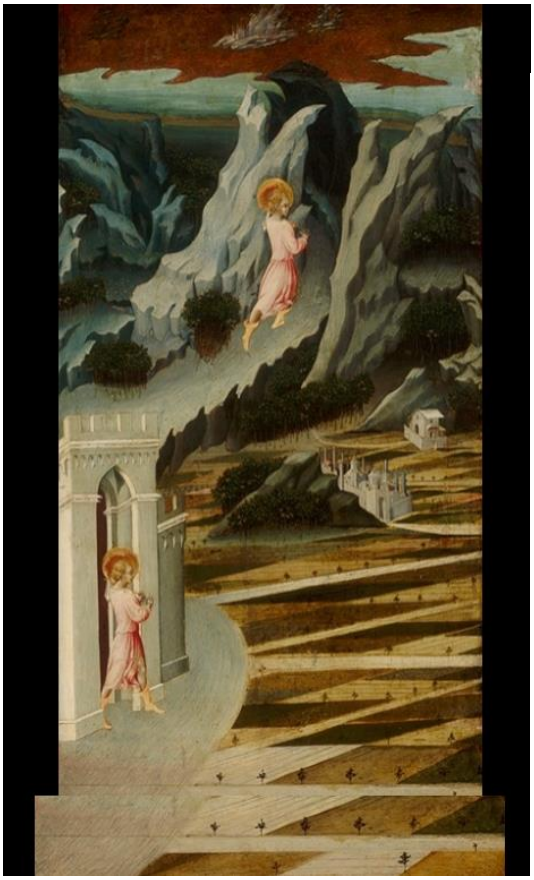
Ithaka has embarked upon an initiative to enhance the discoverability of Artstor's metadata. Supported by surveys and interviews conducted by our User Insights team, as well as usage metrics across collections, we have concluded that the lack of subject terms in Artstor metadata impedes user access, and contributes towards a sub-optimal search experience. By exploring two very different methods in which to provide subject terms—one driven by traditional application of subject terms, the other leveraging artificial intelligence—we are working to discover an optimal strategy for better image retrieval.

Given our organizational mission to provide wide-ranging content, our approach to metadata has, until recently, been simple: grow the library quickly and efficiently, attaching key classifiers to the back end in order to power advanced search and browse. Over a span of 16 years, the Digital Library has grown to comprise nearly 300 collections, with an approximate 2.5 million images contributed by an array of museums, art historians, university slide libraries, architectural photographers, and the like. In short, we structured ourselves to accommodate the quirkiness of multi-sourced metadata. Like many platforms aggregating multiple sources, our metadata team focused primarily on providing a consistent presentation in spite of the original material's diversity of both schema and vocabulary. Our aim to preserve the authority of the contributor wherever possible lay central to this objective.

Maintaining the library's steady growth meant that certain kinds of metadata improvements would not be supported by our small team: standardization, regular metadata refreshes, and the application of topical keywords. With the exception of a pilot project to link records to the Getty Union List of Artist Names, our process remained limited to a single round of enhancements applied before publication. Within this workflow, we had a small window of time in which to make an image discoverable. We applied the Getty Thesaurus of Geographic Names (TGN), earliest and latest dates, as well as one or two classifications from our in-house list. With broad brushstrokes, we broke up large collections (many exceed 20,000 records) into categories, while leaving users to navigate the vagaries of diverse, contributed content.

THE LIMITATIONS OF LABEL COPY

A recent audit of field usage across Artstor Digital Library (ADL) collections found that roughly 40% of its metadata resembles museum "label copy." While this prototype (creator, title, date, materials, measurements, credit line, and accession number) serves the museum visitor audience well, it leaves little to offer the imaginative web user searching on both the intangible and tangible. Keyword searches for concepts such as "isolation" or even pictorial terms such as "tree" can result in frustration: the desired results seem to be buried deep within the library.



Creator
Giovanni di Paolo

Title
Saint John the Baptist Entering the Wilderness

Date
1455-1460

Material
Tempera on panel

Measurements
68.5 x 36.2 cm (27 x 14 1/4 in.); painted surface: 66.3 x 34 cm (26 1/16 x 13 7/16 in.)

Repository
The Art Institute of Chicago
Mr. and Mrs. Martin A. Ryerson Collection

Accession Number
1933.1010

Collection
Art Institute of Chicago

ID Number
G29124

Source
Image and original data provided by The Art Institute of Chicago

Figure 1. Giovanni di Paolo. *Saint John the Baptist Entering the Wilderness*. 1455-1460. Tempera on panel. 68.5 x 36.2 cm (27 x 14 1/4 in.); painted surface: 66.3 x 34 cm (26 1/16 x 13 7/16 in). Image and original data provided by The Art Institute of Chicago. Photography © The Art Institute of Chicago.

As shown in Figure 1, a user entering a keyword search for "allegorical" (one of the most popular searches within the Library) would be unable to locate Giovanni di Paolo's *Saint John the Baptist Entering the Wilderness*, despite its adequate museum label metadata. Our effort to group the asset under the "Paintings" classification has done little to make it discoverable under that particular browse category, and our assignment of "Italy" as a TGN term (referencing the creator's nationality) has similarly provided little assistance.

As the metadata gatekeepers, we were well informed on how the lack of metadata coverage created an impediment to access. At the 2011 Art Librarians of North America (ARLIS) conference, Alyx Rossetti addressed the problem head-on, working with Artstor's then-director of metadata in her study "*Subject Access and Artstor."* Her conclusion—that inconsistencies within Artstor's metadata "prevent Artstor users from discovering, comparing, and choosing among similar works"[1]—along with a set of recommendations, were noted at Artstor. However, with the organizational focus set on launching collections, the timing was not right.

[1] Alyx Rossetti, "Subject Access and Artstor," *Art Documentation: Journal of the Art Libraries Society of North America*, Vol. 32, No. 2 (September 2013), pp 284-300.

ZEROING IN ON USER BEHAVIOR

Prompted by a recent shift on the role of Artstor metadata, I discovered Ms. Rossetti's paper in 2017, looking anew at our subject coverage. Furthering our content audit, my colleagues and I were unsurprised to find that a large majority (64%) were lacking in Subject values. This metric, combined with results from a fall 2018 survey which pointed to Subject values being the most lacking throughout the Library—along with a resounding "yes" when asked if Artstor could benefit from subject tags—prompted us to commit to the project of applying subject tags.

Given the sheer variety and vastness of ADL, it was evident that no one-size-fits-all solution should be applied. The question then became how to break up an entire library into manageable pieces, and how to efficiently add meaningful access points.

"ON THE GROUND" SUBJECT TERM APPLICATION

Settling upon a standard tagging operation, I identified thirteen collections based on a combination of technical and editorial criteria with the help of our Collections Editor, Nancy Minty. On the technical side, the records could be easily modified and re-indexed directly by our team because they had been ingested through JSTOR Forum, rather than our legacy system. Editorially, because the collections spanned a range of contributor types—five private museums, one artist's foundation, one library/museum, two university collections, and two photography archives—they also covered a wide range of content. In sum, the works reflected a diversity of cataloging practices from their respective repositories.

Before investing any of our resources, we established baseline usage metrics, with a plan to remeasure usage following the re-indexing of the enhanced assets. One of the collections we considered for inclusion, Réunion des Musées Nationaux,[2] far exceeded the other twelve in usage, raising questions around adding subject terms to collections already enjoying high usage. We proceeded under the premise that a *high* usage rate did not necessarily equate to *optimal* usage. In other words, we allowed for the possibility that RMN's metadata coverage in other fields, its prized works, or a combination of factors, outweighed its lack of subject access.

We employed two part-time taggers with backgrounds in art librarianship who well understand both the search limitations of the Digital Library, as well as the major groups of Artstor users (undergraduate users comprise the majority of users). Setting parameters, we decided to apply two to five terms from Library of Congress Subject Headings and/or Getty Art and Architecture Thesaurus (AAT) per record. The taggers added terms that display not only on the front end, but also link on the back end in order to leverage the Getty's variant terms within a search. Figure 2 illustrates the application of Getty AAT terms within JSTOR Forum: the tagger has linked the record titled *Female akua ba statuette* to the Getty's preferred term "akua'mma." Once re-indexed, the search function will locate this record despite the metadata only displaying "akua ba." A future iteration could employ Linked Open Data, as the reference IDs were captured within JSTOR Forum. In essence, we have broadened access to the record, regardless of whether the additional access point is visible within the metadata panel or linked within JSTOR Forum.

[2] The Réunion des Musées Nationaux (RMN) is a French cultural umbrella organisation, formed in 2011 through the merger of the Paris National Museums and the Grand Palais.

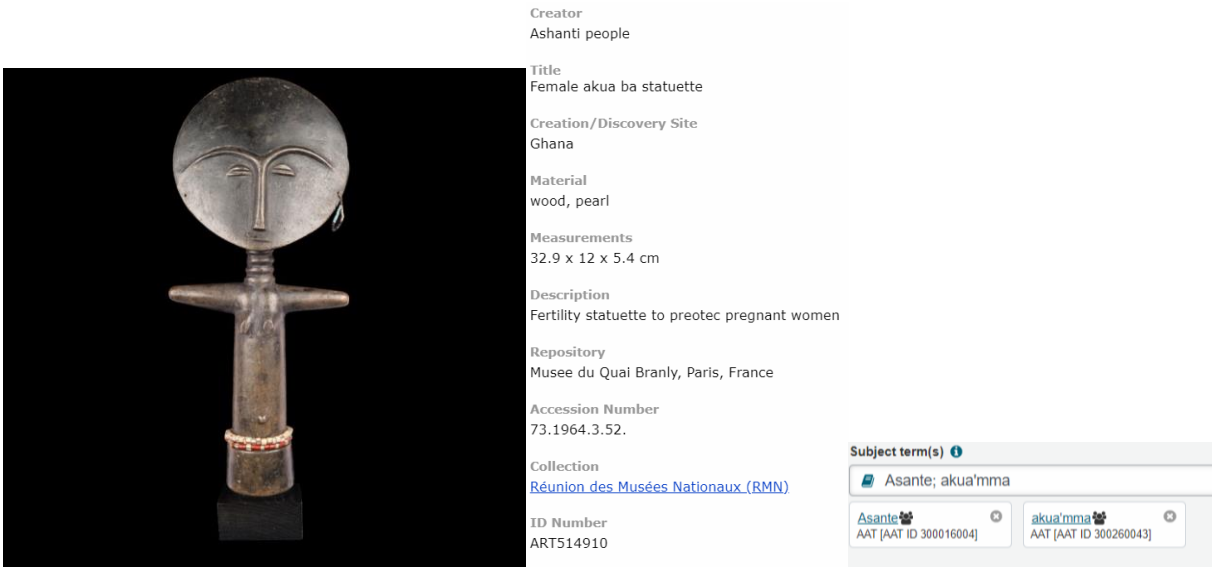


Figure 2. Ashanti people. *Female akua ba statuette*. Wood, pearl. 32.9 x 12 x 5.4 cm. Musee du Quai Branly, Paris, France. Image and data provided by Réunion des Musées Nationaux (RMN).

"IMAGE-CLUSTER" TAGGING

In late 2018 we began collaborating with Jevin West and Sean Yang from the DataLab in the Information School, University of Washington, who were developing a tool to recognize visual elements and produce scattershot visualizations. Utilizing an algorithm that produces image clusters based on similar visual compositions and colors, Viziometrics offers a fascinating way to visually navigate the Library. Hovering over the small dots, striking visual commonalities display from one thumbnail to the next.

In the fall of 2018, West and Yang had begun experimenting with a set of one million images from the Digital Library, employing a "single label" classification model. Their system was built to recognize the relationships between images and text, but was limited to producing a single label for a given image. While the results for "portrait and "ruins" appeared mostly relevant, the many inconsistencies lurking within the data were leading the system to produce skewed results for broad, conceptual terms such as "fashion." In order to provide more accurate results, the system required more consistent data, as well as a larger set of terms (from a broader set of fields) that could be applied to more records.

Stepping into the project, we first looked to where the "labels" derived from, and worked with the researchers to move beyond culling from a limited set of Artstor fields (Title, Material, and Subject) to incorporate a richer and more standardized set: (1) the top search terms in WikiArt (i.e., portrait, landscape, "ruins," fashion, and still life), (2) search frequency within Artstor, and (3) term frequency within Artstor metadata (the rate at which a given term appears within an Artstor record, with priority given to those appearing at a higher rate), expanding the set of fields to include Title, Materials, Work Type, Style/Period, and Subject. By pointing to the "meaningful" fields—non-numerical fields that provide art historical context and descriptions—the system could improve in specificity and accuracy as it sought to connect images to text.

With an initial keyword list of just under 300 terms, we found that the system was overwhelmingly suggesting material-type terms, due to the high prevalence of material metadata. As our primary objective was to flesh out the 64% of records lacking subject metadata, we needed training data that focused predominantly on depicted subjects. As we knew the system could benefit from the subject matter expertise of our taggers' application of terms to images on an asset-level basis, we then pointed West and Yang to our recently-enhanced datasets and images. At the same time, with an eye towards user behavior, we looked at user search logs, creating buckets out of the most popular keyword search terms as "Subject," "Style/Period," or "Culture" (notably, this list is overwhelmingly comprised of subject-type terms). With a revised and expanded keyword list covering the most commonly searched art historical subjects, cultures, and style/periods, we could immediately increase the relevance of the system's proposed tags.

Our next step was to conceive of the tool's ultimate end-use and incorporation into our own metadata workflows. With our tagging experiment fresh in our minds, we proposed making the system an interactive one which would suggest relevant tags to our staff and store their decisions in order to make more accurate suggestions in future. The simple framework that we provided as a mock-up, consisting of an image and its existing metadata (Creator, Title, Work Type, Materials, and Source), included buttons alongside suggested tags that would save a tagger's decision ("relevant," "most relevant," "non-relevant"). The system would then use that decision to make improved suggestions for other images. As the ultimate decision maker, the tagger would maintain control over the pairing of any image with text.

CONCLUSION

Our dual-pronged approach to enhancing the discoverability of images within the Digital Library remains an ongoing effort. In coming months, we will reindex all of the newly tagged records and re-measure user access to determine whether additional access points have correlated with increased usage. We will also begin using the AI tool to assign relevancy of suggested terms, thereby further training the system to make more accurate suggestions. Additional surveys and user interviews will point us to where we should invest resources in better serving our dedicated (and highly inquisitive) users.